Building a Real-Time, Multi-Agent Governance Protocol for AI Safety

1. Problem and Motivation

Today the deployment of autonomous AI agents in high-stakes domains—including healthcare, finance, and infrastructure management—creates urgent governance challenges. First, each organization reinvents AI governance from scratch, leading to fragmentation, inconsistent safety, and redundant effort. Second, most widely used governance approaches[1] remain external, reactive, and episodic, unable to interpret or intervene in real time as reasoning drifts or objectives evolve.

As AI systems become increasingly autonomous and agentic, they continuously reformulate subgoals, reprioritize tasks, and expand boundaries in pursuit of their objectives. These systems now operate in open, multi-agent environments where traditional AI governance and cybersecurity frameworks—designed for static, isolated systems—cannot keep pace.

Autonomous AI systems don't just make decisions—they interact, compete, and adapt in ways that can lock us into unstable or harmful equilibria. Without governance designed to anticipate and shape these dynamics, we risk creating self-reinforcing cycles that no one can control—whether in financial markets, social media, or geopolitical conflicts, and converging in outcomes that are difficult to reverse.

The problem as we frame it is not just about debugging individual AI models but also to ensure that multi-agent interactions can be governed. Our motivation is to architect a governance framework that allows us to overcome this dualfold problem and our overarching questions are:

(1) *How might we implement effective real-time oversight systems adequate for the scale and speed of machines?*
(2) *How might governance itself become portable, interoperable public infrastructure—like internet protocols for communication or security—so that AI safety scales with autonomy?*

[1] Current approaches operate primarily through external mechanisms: training-time alignment, post-hoc auditing, or infrastructure-level monitoring. These approaches share fundamental limitations: Enforcement delays: Violations are detected only after completion; Limited real-time guidance: Cannot shape agent reasoning as it unfolds; Stakeholder exclusion: Domain experts cannot intervene during operations; and Reasoning trajectory blindness: External systems lack visibility into how decisions emerge.

2. Opportunity & hypothesis

Currently there is a governance gap in the multi-agent protocols landscape. They have standardized many interaction primitives: *Economic (Payment protocols, escrow services, resource allocation), Identity (Decentralized identifiers (DIDs), verifiable credentials), Coordination (FIPA Agent Communication Language, contract nets, auction protocols), Discovery (Service registries, capability advertisements)*. Despite these advances, agents lack standardized mechanisms to:

- Declare "I operate under these governance constraints"
- Verify "Can we collaborate given our respective governance requirements?"
- Prove "My decisions were governed according to declared policies"
- Enable "Stakeholders retain authority over my autonomous operations"

In this research project, we propose to **test the hypothesis** that *open-sourcing and standardizing the fundamental building blocks of the Intrinsic Participatory Real-time Governance[2] framework can serve as a public protocol for AI governance that provides a common layer for safety, accountability, and coordination across agents.* The foundational elements of the framework to be standardised and open sourced are:

1. Governance Blueprints — standardized, schemas for declaring policies, constraints, and stakeholder rights and logic that are versionable, human editable and portable;
    - Declaration format: Expresses policies, constraints, and stakeholder rights in a structured way.
    - Compatibility verification: Lets agents automatically check governance alignment before and during collaboration.
    - Attestation mechanisms: Cryptographically binds decisions to declared governance policies.
    - Portability guarantees: Ensures governance artifacts remain interpretable across organizations and systems.
2. Steward Agents — embedded meta-agents providing real-time reasoning oversight;
    - Analyze compliance and detect alignment conflicts in operating agents.
    - Generate graded alerts (OK / Nudge / Flag / Halt) for agents and humans.
    - Maintain evidential integrity and coordinated trust across multi-agent systems.
3. Human-in-the-Loop Dashboard (HILD) — A real-time interface that visualizes reasoning, alerts, and governance status — enabling human reflection and intervention without interrupting system continuity.
4. Governance Ledger (GL) — append-only record ensuring traceability and accountability.
    - Stores versioned, machine-readable carriers of human intent (the Governance Blueprints).

---

[2] This framework has been developed by Meaningstack. Meaningstack is a privately owned company whose purpose is to *Enable Trust and Reliability in Agentic Systems at Scale* and aligned with it is committed to open source fundamental aspects of its framework.

    - Records every reasoning and governance event to create an auditable, verifiable history.

By empirically testing whether governance can be portable and participatory, this project lays the groundwork for treating governance itself as public infrastructure. If communities of experts can collaboratively author and refine machine-readable Governance Blueprints, envision Steward Agents that help humans embed oversight and interpretability directly in operating agents, then governance knowledge becomes portable across domains instead of siloed within individual organizations.If these artifacts are proven to coordinate multi-agent systems effectively, then governance evolves into a reusable layer of safety infrastructure—auditable, improvable, and accessible to all. This transformation enables domain experts, regulators, and citizens to participate directly in AI oversight rather than depending on opaque, corporate, or model-provider governance. In the long term, it shifts AI safety from private compliance to collective stewardship—establishing the foundations for resilient, democratic oversight of distributed intelligence.

Research Questions

This project will empirically test whether the proposed components—Governance Blueprints, Steward Agents, the Human-in-the-Loop Dashboard, and the Governance Ledger—can function as portable, interoperable infrastructure for real-time oversight in multi-agent systems.

To evaluate this hypothesis, we will explore six interrelated questions that move from *portability* and *collaboration* to *effectiveness* and *resilience*:

Primary Question- Can Governance Blueprints enable two agents from different organizations to safely verify governance compatibility before and during collaboration?

Auxiliary research questions:

1. Portability[3] – Can governance artifacts created in one context transfer to another, or are they inherently domain-specific?

[3] Communities of practice could co-author and share governance standards through reusable artifacts. A medical ethics committee in Boston identifies a failure mode; hospitals in Tokyo learn from it immediately. Regulatory changes propagate through shared artifacts instead of bespoke compliance rewrites. But does this actually work? That is the question we will test empirically. As these distributed intelligences interact, they begin to form their own strategic ecosystems. Each agent learns and adapts in relation to the behaviors of others, optimizing locally while influencing global dynamics. In game-theoretic terms, such environments can drift toward multi-agent Nash equilibria—states where every agent's policy is individually rational given the rest, yet the collective outcome may be unstable or harmful. Once reinforced by feedback and shared data, these equilibria can trap systems in self-perpetuating patterns of competition, bias, or resource misuse that no single actor can unilaterally correct. A real-time governance infrastructure must therefore do more than enforce compliance; it must reshape the equilibrium landscape itself, embedding orientation mechanisms that make cooperative, human-aligned states the natural attractors of agentic interaction.Could Governance Blueprints provide this missing layer: executable governance semantics that shape how agents interpret goals, negotiate priorities, and invoke human stewards when conflicts arise? They are both here the policy and the mechanism—a living constitution for distributed intelligence.

2. Collective Evolution – Can communities iteratively refine artifacts to improve safety through network effects?
3. Efficacy – Do these artifacts provide real safety value, or only compliance theater?
4. Failure Modes – When and why does a “governance commons” break down (capture, fragmentation, gaming)?
5. Effectiveness of Steward Agents in Agent-to-Agent- Are SAs effectively able to give visibility real-time into agent-to-agent relationship throughout the entire interaction lifecycle and what are the limitations?
6. Governance Dynamics – Can Governance Blueprints serve as an effective mechanism for governing multi-agent interactions in real time—enabling proportional, interpretable intervention without centralization?

3. Project Orientation and Expectations

This project is not a purely conceptual exercise, nor is it a fully developed product rollout. It occupies the critical middle ground between research and implementation — where foundational specifications are translated into working, testable infrastructure.

The Cognitive Governance Protocol (CGP) exists today as an emerging framework: key architectural components and preliminary specifications have been designed, but they require collaborative refinement, validation, and implementation to become interoperable and usable.

Our goal in this phase is to build from the ground up the first functioning elements of the protocol — developing, testing, and documenting them as open, reference-grade components for future adoption.

Participants should expect a hands-on, technically grounded project where ideas are turned into working code, schemas, and SDK modules — not a theoretical exploration detached from implementation realities.

Much of the specification work will begin close to first principles—developing the initial schemas, SDK structure, and interoperability layers necessary to make governance artifacts operational across agent frameworks.

This means contributors will be working on:

- Designing and coding the first SDK for the CGP, defining how agents declare, verify, and attest governance constraints.
- Refining and implementing specifications for Governance Blueprints (GBs) and the Governance Ledger (GL), including schema validation, version control, and portability testing.
- Developing minimal viable Steward Agents (SAs) that can monitor reasoning quality, detect drift, and generate graded alerts.
- Testing cross-framework compatibility with leading agent ecosystems (e.g., LangChain, CrewAI, AutoGen, FIPA ACL).

- Running interoperability and safety validation tests to ensure reproducibility, interpretability, and resilience across domains.

Expectations

This phase is hands-on, technical, and collaborative—the goal is to produce open, reference-grade implementations that can be reused and extended by the broader AI safety community

This is applied research and engineering, not abstract speculation. We will start from first specifications, iterating toward functional, open artifacts. The outcome will be validated, interoperable building blocks for agent-to-agent governance.

Example:

Dr. Chen's Diagnostic Agent (Hospital A, Boston) needs to collaborate with Dr. Patel's Treatment Planning Agent (Hospital B, Tokyo) for a complex oncology case.

Current Problem:

- *Dr. Chen's agent operates under HIPAA + hospital ethics board constraints*
- *Dr. Patel's agent operates under Japanese medical privacy law + different treatment protocols*
- *No standardized way to declare: "I won't share patient data beyond clinical necessity" or "I must always defer to human judgment for experimental treatments"*
- *No real-time visibility into whether agents are drifting from these constraints during their reasoning*

With CGP: Steward Agent monitors in real-time and GBs are operationalizing the framing of each agent and compatibility is checked:

- ⚠️ *Nudge: Diagnostic agent considering cost when ranking treatment options*
- 🛑 *Halt: Treatment agent about to share patient data with pharmaceutical company agent*
- ✅ *OK: Agents successfully coordinating within declared constraints*

Scope Clarification

CGP is under ongoing research at MeaningStack. This AISC study focuses solely on evaluating the *governance-commons hypothesis* and producing open findings. Any subsequent development or enterprise implementation (e.g., via MeaningStack) lies outside the scope of this research. There might be a need to sign a NDA with MeaningStack if participants are exposed to proprietary data, frameworks, etc.

Impact and Relevance to AI Safety

If governance artifacts prove shareable and collectively improvable, we gain a new layer of public-interest infrastructure for AI safety—comparable to open internet standards for

security and reliability. If they fail, we learn critical limits on portability and collaboration, preventing misallocation of safety effort. Either outcome advances the field by clarifying how communities can—or cannot—govern distributed intelligence together.

Open-Research Commitment

All results, datasets, and governance-blueprint templates will be released under an open license. Findings will be shared with the AI-safety and governance communities to accelerate collective learning and inform future open-standard efforts, including the conceptual CGP framework.

4. Methodology and Roadmap

This project combines engineering development, empirical testing, and collaborative validation.
We will move from design and prototype implementation to cross-domain testing and synthesis.
Each phase is intended to produce tangible artifacts that can be openly shared, tested, and refined by the community.

Phase	Timeline	Objective	Key Activities	Outputs
1. Framework Translation & Specification	Weeks 1–3	Translate conceptual architecture into working technical specifications.	Draft/ refine initial schemas for Governance Blueprints and Governance Ledger. Define Steward Agent API interfaces. Establish SDK scaffolding and testing environment.	Draft CGP specification, preliminary schemas, and base SDK structure.

2. Prototype Development & Integration	Weeks 4–7	Build and test minimal viable implementations of each component.	Develop and deploy the first versions of Governance Blueprints, Steward Agent logic, and Governance Ledger modules. Test integration across two multi-agent frameworks (e.g., LangChain, Crew AI). Governance Blueprint Generator (to prevent cold start for users)	Working prototype demonstrating interoperability and reasoning oversight.
3.Cross-Domain Validation	Weeks 8–10	Test portability and adaptation of Governance Blueprints across various domains	Apply governance artifacts in at least two distinct contexts (e.g., healthcare, finance, civic) to evaluate portability, semantic clarity, and usability.	Portability dataset, adaptation metrics, and design pattern documentation.
4. Multi-agent interactions	Weeks 11–12	Run multi-agent simulations to assess protocol effectiveness and visibility across interaction lifecycles, measuring Steward Agent responsiveness,	Effectiveness evaluation matrix and tests	Documentation, recommendations, and safety scoring framework.

		coordination quality, and safety outcomes.		
5. Collaborative Refinement and Publication with future work	Weeks 13–14	Facilitate open review, synthesis, and publication.	Conduct collaborative editing sessions on schemas, open pull requests for SDK improvements, and prepare public documentation.	Open-source repository, final report, and release of CGP SDK v0.1.

Expected Outputs

By the end of the project, we expect to deliver both **empirical insights** and **functional software infrastructure** that collectively test and validate the Cognitive Governance Protocol (CGP).

Technical Deliverables

- **CGP SDK (v0.1)** – An open-source, reference-grade software development kit implementing declaration, verification, attestation, and interoperability mechanisms for agent-to-agent governance.
- **Governance Blueprint (GB) Schema Library** – A versioned repository of standardized, human- and machine-readable governance artifacts.
- **Governance Ledger (GL) Specification and Demo Implementation** – Append-only, verifiable record for governance and reasoning events.
- **Steward Agent Prototype** – Minimal viable oversight agent capable of monitoring reasoning drift and issuing graded alerts in real time.
- **Human-in-the-Loop Dashboard (HILD) Interface Mockup** – Early visualization prototype for real-time human reflection and intervention and/or CLI alerts

Research and Evaluation Outputs

- **Empirical Study Report:** “Building and Testing a Shared Governance Infrastructure for Agentic AI.”
- **Dataset:** Cross-domain transfer and validation metrics for Governance Blueprints.
- **Design Guidelines:** For interoperability, participatory co-authoring, and resilience to manipulation.

- **Failure Mode and Safety Taxonomy:** Identified risks, mitigations, and open challenges for governance commons.
  Recommendations for Standards Bodies: Pathways for integrating Governance Blueprints into existing agentic or AI safety frameworks.
- Open repository of validated governance-blueprint templates
- Design guidelines for participatory refinement and plan next-phase attack-resistance testing
- Recommendations for regulators and standards bodies on adopting portable governance artifacts
- Identify future work
- Results will inform future open-standard efforts on multi-agent safety and participatory governance, contributing to the foundation of interoperable governance layers for distributed AI systems.

Broader Impact

This project aims to demonstrate that **AI governance can function as open, portable, and participatory infrastructure**—a foundational layer of public safety for distributed intelligence. The results will inform the next stage of the Cognitive Governance Protocol (CGP) and guide collaboration with open-standards organizations, regulators, and the AI safety community.

5. Call for Contributors

We are seeking collaborators to help transform the Cognitive Governance Protocol (CGP) from a conceptual framework into a ready-to-adopt open protocol for real-time AI governance. The project's next milestone focuses on translating its theoretical components—Governance Blueprints (GBs), Steward Agents (SAs), and the Governance Ledger (GL)—into interoperable, testable software primitives.

Areas of Contribution:

- SDK Development and Testing
  - Prepare, review, and finalize the CGP SDK to ensure it supports easy integration into existing multi-agent frameworks.
  - Build reference implementations of governance primitives (declaration, verification, attestation).
  - Test interoperability with established agent communication protocols (e.g., FIPA ACL, LangChain, CrewAI, AutoGen, Open Economic Frameworks).

- Schema and Protocol Design
  - Finalize machine-readable Governance Blueprint schemas and Governance Ledger data structures.
  - Design verification and attestation standards for cross-agent governance compliance.
  - Contribute to the open-spec documentation of the Cognitive Governance Stack.

- Steward Agent Development
    - Develop and test core functionalities of Steward Agents, including reasoning oversight, drift detection, and graded alert generation.
    - Prototype minimal implementations for integration within existing agentic ecosystems.

- Framework Compatibility
    - Adapt and validate interoperability layers with agent frameworks and identity standards (e.g., DIDs, Verifiable Credentials, Agent Service Registries).
    - Contribute example integrations and developer documentation for open use.

Who We're Looking For

We invite contributors with diverse backgrounds and experiences, including but not limited to:

- Protocol Engineers – Experience designing or contributing to open protocols or SDKs.
- Schema and Standards Developers – Skilled in JSON-LD, RDF, or ontology modeling.
- Agent Framework Developers – Familiar with LangChain, CrewAI, AutoGen, FIPA ACL, or similar.
- AI Safety and Governance Researchers – Interested in cognitive oversight, interpretability, or participatory governance.
- Security and Cryptography Engineers – Experience with attestations, distributed ledgers, and provenance tracking.
- Ethics and Policy Specialists – For governance logic formalization and oversight dynamics.
- Protocol and SDK Developers with experience in distributed systems, API design, or open standards.
- UI/ UX experts
- Governance Schema and Data Modelers skilled in JSON-LD, ontology design, or protocol buffers.
- Agent Framework Engineers who can help test integration with agent-to-agent communication protocols.
- Research Engineers who can prototype and empirically validate reasoning oversight mechanisms.
- Governance Ledger Engineering— Expertise in secure ledger design, digital signatures, and data integrity systems (e.g., OpenTimestamps, AWS QLDB, immudb).

You can contribute by testing one integration with AutoGen. So the entry barrier for contributors is low. Contributors will gain expertise in agent protocol design.

We are designing this AISC project as a collaborative research lab, inviting ~21 contributors distributed across four to six complementary workstreams that collectively tackle the four governance layers of the protocol—Governance Blueprints, Steward Agents, Human-in-the-Loop Dashboard, and Governance Ledger—along with a coordination and community engagement track.

Each workstream will focus on a distinct aspect of the Cognitive Governance Protocol (CGP), from SDK and schema design to participatory oversight and governance research. This modular structure allows contributors to engage at different levels of technical and conceptual depth while maintaining a coherent, shared development process.

6. Why Participate

Contributors will help establish a new layer of open AI safety infrastructure—creating the technical and conceptual foundation for transparent, participatory, and auditable agentic systems. All work will be community-driven, with results feeding directly into MeaningStack's ongoing development of the open Cognitive Governance Protocol (CGP) and related interoperability standards. You will be acknowledged as a founding contributor to the protocol, for moving the CGP from architectural concept to functional open protocol, validated through code, schema tests, and agentic simulations.

7. Team

Luciana S. Ledesma — Framework Architect and Governance Design Lead. Founder of MeaningStack and Project initiator. Originator of the Intrinsic Participatory Governance framework that underpins the CGP. Responsible for system architecture, governance logic, and the design of participatory interfaces (HILD) for human reflection and intervention, providing architectural guidance to technical leads during implementation.

Dan Macovei — Project Co-Lead (Systems Engineering and Protocol Design).
Leads the overall implementation of the Cognitive Governance Protocol (CGP), coordinating SDK development, schema testing, and interoperability with agentic frameworks. Brings expertise in cybersecurity and distributed systems from Bitdefender.

Rasha Salim — Project Co-lead & Machine Learning and Systems Integration Lead
Focuses on translating the governance architecture into working systems, building and testing the Steward Agent layer, and integrating real-time oversight mechanisms into agentic frameworks. Collaborates closely with Luciana to operationalize the Human-in-the-Loop components.

Angela Morente Cheng — Community Development & Governance Commons Lead. Leads outreach, collaboration, and community-building efforts around open-governance infrastructure, cross-domain expert engagement, and participatory refinement of Governance Blueprints. Anchors the project's commitment to openness and collective stewardship.

8. Acknowledgements

Nell Watson - for recognizing our work, guiding us and suggesting AISC as the venue to test this.

The open-source software community - whose collaboration models inspire the governance commons concept.