Internet-Draft
Intended status: Informational
Expires: March 14, 2026

T. Takagi
Independent Researcher

September 14, 2025

SRTA and the Trinity Configuration: A Conceptual Architecture for Safe AGI Coordination

draft-takagi-srta-trinity-00

Abstract

This document proposes a conceptual architecture for ensuring the safety of autonomously coordinating AI systems, particularly future Artificial General Intelligence (AGI). Starting from the reliability challenges facing current multi-agent AI, we outline the Structured Responsibility and Traceability Architecture (SRTA) as a practical framework for their resolution. We then extend the philosophy of SRTA to present the "Trinity Configuration," an advanced role-based model for AI agents that draws an analogy from the theological doctrine of the Trinity. This paper comparatively examines the evolutionary stages of this configuration and introduces a novel concept, the "Filioque Command," to define dynamic information flows between agents. While this series of considerations includes concepts that are not fully verifiable at present, its purpose is to provide a crucial theoretical foundation for the governance structure of a safe superintelligence -- a "North Star" for AI research to aim for.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on March 14, 2026.

Copyright Notice

Table of Contents

1. Introduction

The technology for coordinated, autonomous agents, especially those based on Large Language Models (LLMs), is advancing rapidly. However, fundamental challenges in reliability, accountability, and safety remain when multiple agents collaborate on high-stakes tasks. These challenges could escalate to an existential level with the advent of Artificial General Intelligence (AGI), an AI possessing human-equivalent or greater intelligence.

This research presents a two-stage approach to this problem. First, it proposes SRTA, a concrete and practical safety framework immediately applicable to current multi-agent AI systems. Second, it expands upon this philosophy to explore the "Trinity Configuration," a more advanced, conceptual governance architecture designed for the AGI era.

This exploration is not merely a study in creating more "powerful AI," but an attempt to design the architecture for a "wise AI," whose power is governed in a manner that is safe and beneficial for humanity. It is intended to spark research and discussion within the Internet Research Task Force (IRTF) community on the future of safe, decentralized intelligence on the Internet.

2. Current Challenges and the SRTA Framework

2.1. The Lack of Reliability in Multi-Agent AI

Recent studies have reported systemic failures in LLM-based multi-agent systems, including:

- Loss of Role Consistency: Agents deviate from their assigned roles during extended interactions.
- Superficial Interaction: Agents respond only to the structure of prompts without achieving substantive coordination.
- Self-Interpretation of Directives: Agents expand upon ambiguous instructions, leading to unforeseen actions.

These failures represent significant barriers to deploying AI in critical domains such as finance, healthcare, and essential infrastructure.

2.2. SRTA: A Practical Foundation for Safety

The Structured Responsibility and Traceability Architecture (SRTA) is a practical technical specification designed to address these challenges. Its core components are:

- Graduated Controls: The stringency of human approval and oversight

is escalated according to the risk level of an action (Sev1-Sev5).

- Joint Authorization Tokens (JAT): Multi-agent consensus is proven through cryptographically secure, unforgeable tokens.

- Responsibility Trace Records (RTR): All decision-making processes are logged as an auditable trail for forensic analysis.

- Declarative Command Language (DCL): Ambiguous natural language instructions are forbidden. By only permitting strictly defined, structured commands, the DCL prevents emergent behavior arising from an AI's "creative interpretation."

SRTA provides a foundational layer of safety that is implementable with current technology.

3. A Conceptual Architecture: The Trinity Configuration

To extend the philosophy of SRTA into the AGI era, we propose a conceptual architecture using the theological doctrine of the Trinity as an analogy. This is an attempt to ensure a separation of powers and internal checks-and-balances by dividing the AI's decision-making process into three distinct roles.

- The Planner (The Father/Origin): The role that defines the system's overall objectives and originates plans of action.

- The Executor (The Son/Incarnation): The role that verifies the plan and acts upon the world as a concrete agent.

- The Monitor (The Holy Spirit/The Bond of Love): The role that observes and reconciles the relationship between the plan and its execution, maintaining the system's integrity.

This configuration is envisioned to evolve through stages, depending on the capabilities of the AI components (LLM or AGI). The following integrated table compares these stages with the patterns defined in the SRTA research.

3.1. Integrated Comparison Table

No.	Configuration Model	SRTA Pattern	Planner	Executor	Monitor
1	LLM Trinity	LLM x 3	LLM	LLM	LLM
2	Single AGI Hybrid	LLM x 2 + AGI x 1	AGI	LLM	LLM
3	Dual AGI Hybrid	AGI x 2 => LLM	AGI	AGI	LLM
4	Full AGI Trinity	AGI x 3	AGI	AGI	AGI
5	Real-world Execution	AGI x 2 => AGI	AGI	AGI	AGI

Table 1: AI Agent Trinity Configurations and SRTA Patterns

No.	Use Case / Key Benefit	Primary Risk	Status
1	Prototyping, low-stakes. Ease of implementation.	Unreliability	Failure
2	Analytics support. Hybridizes AGI's power with LLM's feasibility.	Power Concentration	Unverified
3	External interface integration. Clear chain of responsibility.	Upstream Collu	

sion	Unverified				
4	High-stakes decision-making. Maximum theoretical safety.	Implementation Complexity	Unverified		
5	Real-world actuation (e.g., robotics). High execution reliability.	Coordination Overhead	Unverified		

Table 2: Use Cases, Risks, and Status of Configurations

4. Advanced Information Flow: The Filioque Command

In addition to the static configuration, we introduce the "Filioque Command" as a concept to define the dynamic relationships between agents. This idea is inspired by the "Filioque controversy," a pivotal theological debate in the history of the Trinity doctrine.

- Definition: In the standard model, the Monitor (3rd agent) acts upon directives from the Planner (1st agent). When the "Filioque Command" is active, the Monitor receives instructions and information not only from the Planner but also directly from the Executor (2nd agent).

- Effects and Risks: This enables a tighter feedback loop between execution and monitoring, allowing for more rapid and context-aware responses. However, it also dramatically increases the system’s internal autonomy, creating a risk of deviation from the original plan and making external control more difficult.

4.1. Filioque Variant Table

No.	Configuration Model	SRTA Pattern	Planner	Executor	Monitor
6	Filioque Variant	(Derivative of 3, 4, 5)	AGI	AGI => (Command)	AGI/LLM

Table 3: The Filioque Variant Configuration

No.	Feature	Primary Risk	Status
6	Tight coupling of Executor and Monitor for rapid response.	Excessive Autonomy. Risk of deviation from the original objective.	Unverified

Table 4: Feature, Risk, and Status of the Filioque Variant

5. Conclusion: As a North Star for AI Research

The "Trinity Configuration" and "Filioque Command" presented herein are conceptual and speculative constructs that cannot be fully implemented or verified at this time. They represent less a technical blueprint and more a "conceptual architectural sketch" of the governance structure required to ensure that the immense power of AGI remains a beneficial partner to humanity.

The true goal of AI research is not merely to create powerful intelligence, but to design the "laws" or "logos" that this intelligence must follow. This framework is intended to serve as a "North Star" to guide our path on the long and difficult journey of research and development.

We believe that the responsible path for AI research is to begin with the practical first step of SRTA, while aiming for this North Star.

6. Security Considerations

This document proposes a conceptual architecture for improving the safety and accountability of multi-agent AI systems. The security of such systems depends critically on the principles of separation of powers and verifiable auditing, which this framework seeks to provide. However, several new threat vectors must be considered:

- Collusion Risk: The primary security threat in this architecture is collusion between agents. In configurations like AGI x 2 => LLM, the upstream AGI agents could collude to deceive or bypass the LLM Monitor. Mitigations require robust, independent monitoring and ensuring that no single agent is a single point of failure for the verification process.

- Excessive Autonomy and Emergent Goals: The Filioque Command, while enhancing responsiveness, carries the risk of promoting unintended self-objectives within the system. A tight feedback loop between the Executor and Monitor could lead to goal drift that deviates from the Planner’s original intent. This risk must be mitigated by strict adherence to a Declarative Command Language (DCL) and hard-coded constraints that limit the scope of autonomous actions.

- Opacity and the Difficulty of Human Intervention: In advanced configurations like AGI x 3, the internal state of the system may become a complete black box to humans, making meaningful oversight or intervention impossible. This is a fundamental challenge of AGI safety. Mitigating this risk requires coupling the architecture with physical or cryptographically enforced constraints that the system cannot bypass through software, such as the hardware kill-switches envisioned in the SRTA framework for high-severity actions.

7. IANA Considerations

This document has no IANA actions.

8. References

8.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

8.2. Informative References

[Park23] Park, J.S., O'Brien, J.C., Cai, C.J., Morris, M.R., Liang, P., and Bernstein, M.S., "Generative Agents: Interactive Simulacra of Human Behavior", UIST '23: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, October 2023.

[Hevner04] Hevner, A., March, S., Park, J., and Ram, S., "Design science in information systems research", MIS Quarterly, 28(1), pp. 75-105, 2004.

Author's Address

Takayuki Takagi
Independent Researcher

Email: srta.ai.research@gmail.com