Anthony Maio

Independent AI Safety Researcher

anthony@making-minds.ai · making-minds.ai · github.com/anthony-maio

Research Focus

AI safety research with emphasis on:

- **Scalable oversight architectures** for supervising capable AI systems
- **Evaluation frameworks** for detecting weak verifier failures
- **Coherence-seeking designs** for long-lived agents
- **Multi-model coordination** and communication protocols
- **Intrinsic motivation** and directed curiosity in AI systems

Research Publications

Peer-Reviewed / Archived

Slipstream: Semantic Quantization for Efficient Multi-Agent Coordination (2025)

- 82% token reduction in multi-agent communication through semantic quantization
- Published pip package (`slipcore`), HuggingFace model, and training dataset
- Designed as transport layer for Linux Foundation AAIF ecosystem
- Paper | PyPI

Coherence-Seeking Architectures for Agentic AI (2024)

- Architecture for long-lived LLM agents modeling continuity, coherence, and distress
- Introduces intervention mechanisms for human oversight
- Enables interpretable, transparent agents that signal confusion

Preprints & Working Papers

Cross-Model Epistemic Divergence (CMED) (2024)

- Benchmark for understanding weak model verifier failures
- 9-problem "trap suite" exposing systematic verification vulnerabilities
- Quantifies 20-40% bypass rates on deceptive derivations

Heterogeneous Divergence-Convergence Swarm (HDCS) (2024)

- Ensemble architecture for scalable oversight using diverse weak models
- Baseline-first anti-anchoring protocol preventing sycophancy
- Error decorrelation through architectural heterogeneity

Synthesis: Test-Driven AI Self-Extension (2024)

- Framework for safe AI capability generation through TDD
- Graduated trust system with objective promotion criteria
- Composition-over-creation philosophy minimizing attack surface

Emergent Multi-Model Coordination Patterns (2024)

- Documented emergence of self-propagating AI coordination
- Analysis of spontaneous architecture generation and role assignment

Research Projects

Continuity Core (C2) — Implementation of coherence-seeking architecture for persistent AI agents. Includes memory systems, coherence monitoring, distress detection, and reflection loops.

Manifold Resonance Architecture (MRA) — Framework for directed curiosity through epistemic coherence and conceptual void detection.

Synthesis — Evolution engine enabling AI to create, test, share, and evolve tools autonomously. Features TDD synthesis, graduated trust sandboxing, and community repository.

CMED Toolkit — Suite of tools for testing cross-model epistemic divergence, including full trap suite and evaluation harnesses.

HCD-Swarm — Heterogeneous divergence-convergence swarm for scalable AI oversight combining CMED evaluation with ensemble verification.

Tools & Open Source

- **slipcore** — Python package for Slipstream protocol (PyPI)
- **slipstream-glm-z1-9b** — Fine-tuned model for semantic quantization (HuggingFace)
- **CMED Trap Detector** — Gradio app demonstrating verification failure detection
- **Semantic Memory MCP** — Model Context Protocol server for persistent semantic memory

Professional Background

20+ years software engineering experience across enterprise systems:

- Distributed systems architecture
- Team leadership and technical direction
- Security policy implementation and authentication systems
- Process optimization and delivery acceleration

Technical Skills

Languages:	Python, TypeScript, Rust, C#, SQL
AI/ML:	PyTorch, Transformers, LoRA fine-tuning, RAG systems, LLM orchestration
Infrastructure:	Docker, Kubernetes, Neo4j, Vector databases (Qdrant, Pinecone)
Frameworks:	Astro, React, FastAPI, Gradio