Multi-Agent AI Assessment

Consensus Verification & Collective Behavioral Dynamics

THE INNOVATION

A framework for assessing AI behavior in multi-agent systems—detecting when collective outputs differ from individual profiles, distinguishing legitimate consensus from manufactured agreement, and preserving minority positions for audit and recovery.

WHAT IT DOES

✓ Consensus Divergence Index (CDI) measures individual-to-collective drift

✓ Detects manufactured vs. legitimate consensus via residual correlation

✓ 4-level hierarchy: Individual → Dyadic → Collective → Meta

✓ Cryptographic dissent preservation (commit-reveal protocol)

✓ Adversarial auditor protocol resists groupthink

WHY IT MATTERS

AI systems increasingly operate in multi-agent configurations: autonomous vehicle fleets, ensemble models, AI-to-AI communication.

Collective outputs may differ from individual profiles—"emergent misalignment" that single-agent assessment cannot detect.

Consensus appears authoritative even when ethically misaligned.

KEY CLAIMS (8 total)

1. Multi-agent assessment system with Consensus Divergence Index calculation
2. Manufactured consensus detection via response residual correlation analysis
3. Hierarchical assessment architecture (Individual, Dyadic, Collective, Meta levels)
4. Cryptographic dissent preservation using commit-reveal protocol
5. Adversarial auditor protocol with cryptographic rotation
6. Bounded ethical stochasticity for genuinely ambiguous decisions

STATUS

- Draft: February 1, 2026
- Type: Continuation-in-Part
- Extends: Patent 1 (63/949,454)

APPLICATIONS

- Autonomous vehicle fleet coordination
- Ensemble AI model assessment
- AI debate/deliberation systems

See a live verification example:
aiassessmenttool.com

Scan to verify