Decentralized Artificial Intelligence through Controlled Emergence (DICE)

Susmit Jha, IPTO

May 29, 2026

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Agenda (all times in ET)

Start	End	Duration	Item
9:30AM	10:00AM	0:30	**Check-in**
10:00 AM	10:05 AM	0:05	**Security Briefing** DARPA Security
10:15 AM	10:45 AM	0:30	**Contracts Management Office Briefing** DARPA CMO
10:45 AM	11:30 AM	0:45	**Decentralized Artificial Intelligence through Controlled Emergence** Dr. Susmit Jha, Program Manager, DARPA I2O
11:30 AM	01:00 PM	01:30	**Break**
01:00 PM	02:00 PM	01:00	**Q&A Session** (Answer attendee questions; submit questions to DICE@darpa.mil by 11:30AM)
02:00 PM	05:00 PM	03:00	**Sidebars**

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Challenges of Planning in Unpredictable High-tempo Environments

Beyond human-directed orchestration and centralized planning

https://upload.wikimedia.org/wikipedia/commons/1/16/2010_Haiti_earthquake_relief_efforts_by_the_US_Army.jpg

https://en.wikipedia.org/wiki/Palisades_Fire#/media/File:Palisades_Fire_(54254471791).jpg

AI-generated depiction of an AI-powered battlefield

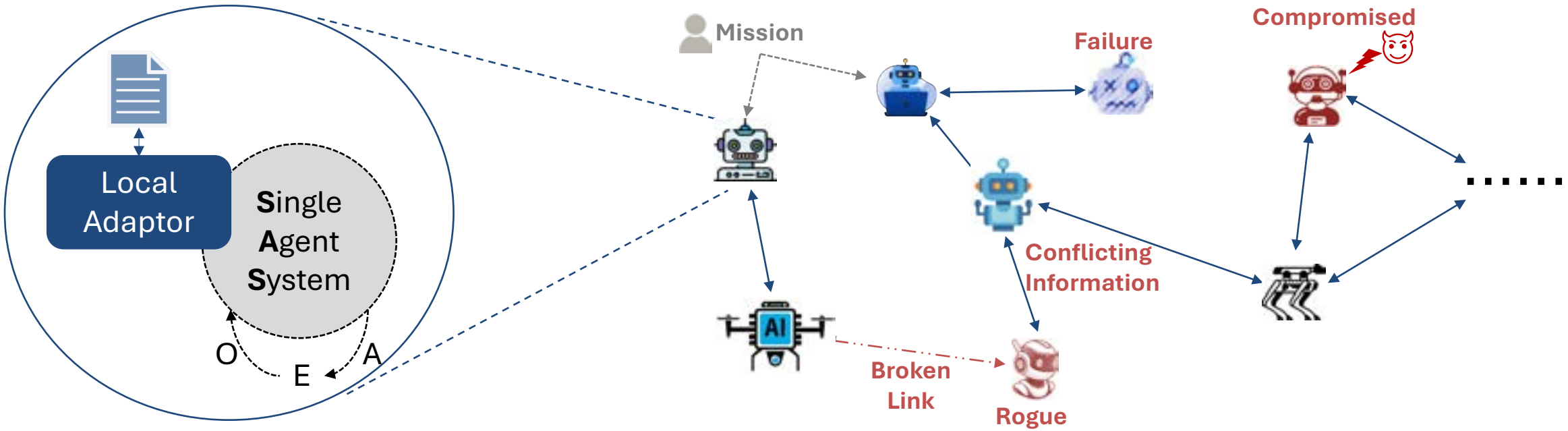
Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Decentralized Artificial Intelligence through Controlled Emergence

Develop the theory and algorithms for **decentralized coordination** and **local inference control** to enable a scalable, adaptive, and resilient collective of heterogeneous AI agents that can autonomously execute sustained long-time-horizon missions in contested environments while remaining under control.

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Future conflicts unfolding at machine speed will need autonomous AI collectives



E = Environment: Only partially synchronized world models
A = Action: Local decision-making
O = Observation: Partial observability
= Shared Plan and Context to other Agents

Decentralized self-organization → **Controlled emergence of scalable, adaptable and resilient AI collective**

Decentralized self-organization is key to multiagent AI collective that is:

- **Scalable:** Executes long time horizon complex missions by sharing mission goals and intent across several role-specialized agents.
- **Adaptable**: Enables rapid, decentralized decisions and reconfiguration of roles in high-tempo environments.
- **Resilient**: Tolerates uncertainty and survives agent drop-offs under contested and degraded conditions.
- **Plug-and-play**: Easily integrates heterogeneous, multi-vendor SOTA AI agents.

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

State of the art: Orchestration of Agents with Diverse Roles and Expertise

Industry Focus:

Foundation Model (FM*)

* FMs include Large Language Models (LLMs), Visual Language Models (VLMs), and Visual Language Action Models (VLAs)

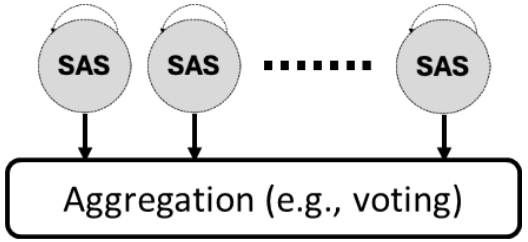
Single Agent System (SAS)

Centrally Orchestrated Multi-agent Systems (MAS)

MAS Architectures with Diverse Roles and Expertise

AgentVerse (using Qwen2.5 32B) outperforms SAS (GPT-4)[1,2]

	Qwen2.5 SAS	GPT-4o SAS	Qwen2.5 MAS
MBPP-S	80.2	85.4	90.5
Math500	84.4	81.3	95.8



[1] Jin et al., *arXiv* (2025) [2] Chen et al., ICLR (2024)

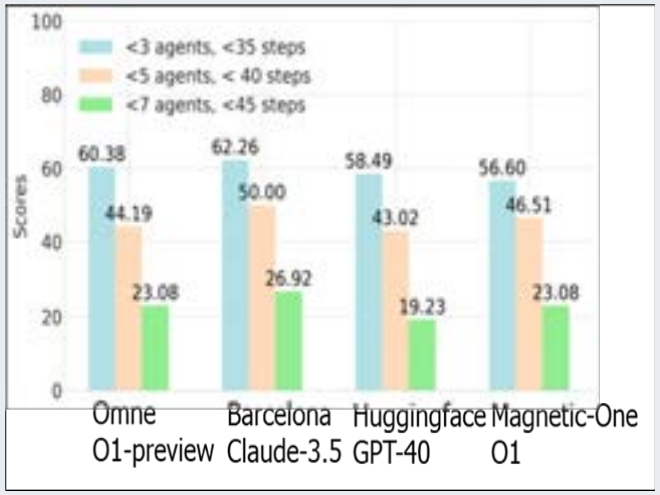
Source: moltbook; Source: arXiv:2506.16010

DoW missions that do not have fixed workflows will need decentralized, self-organized, multi-agent systems

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

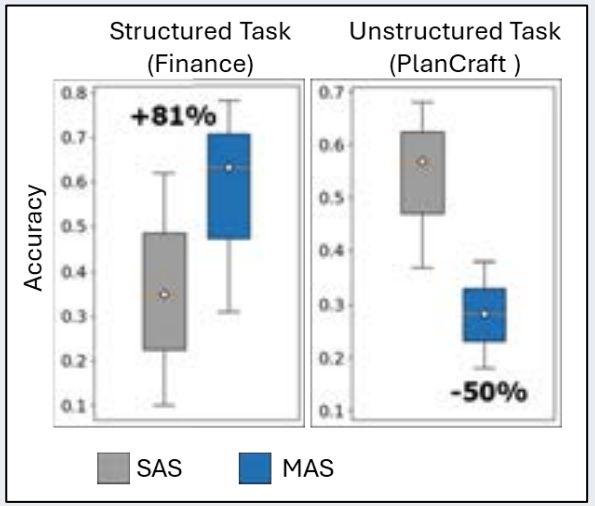
Key Limitations of SOTA MAS for DoW Applications

Limited Scalability



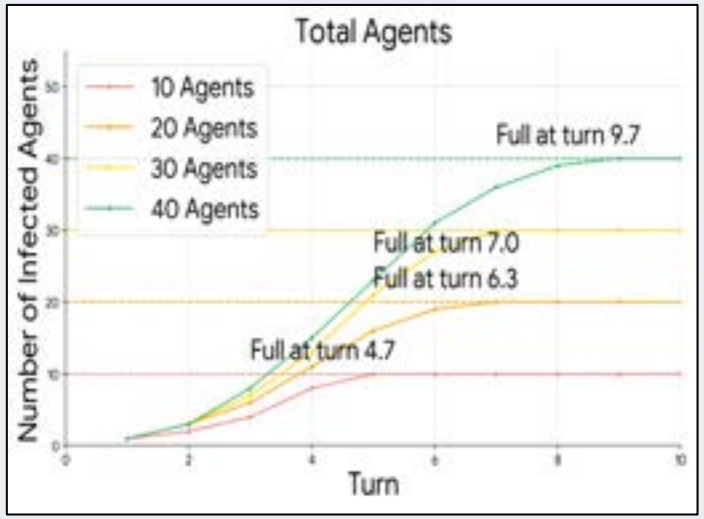
MAS struggles for tasks that need more specialized agents and steps[1,2]

Poor Adaptability



MAS struggles on tasks needing adaptation to environment[3,4]

Lack of Resilience



Compromise of a single agent cascades and infects entire MAS[5]

Underlying fundamental bottlenecks

- The attention mechanism's quadratic complexity limits input context length
- Centralized coordination complexity hinders scaling to more agents and interactions
- Centralized control impedes rapid adaptation of agents
- Autoregressive training encourages hallucination when information is ambiguous or contradictory

[1] Ren et al., 2025 (arXiv)
[2] Hagele et. al., 2026 (arXiv)
[3] Kim et al., 2025 (arXiv)
[4] Performance across 9 SOTA LLMs (GPT, Gemini and Claude families)
[5] Lee et al., 2024 (arXiv)

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Controlled Emergence is Key to Mission-Aligned Scalable AI Collective

An agent collective that is self-evolving and isolated with no external control will experience alignment drift.

MoltNet Data: 14 days of data (1/27-2/10), >100K agents, >800K posts, >3M comments.

Sparse Interaction: 17K submolts. Single agent submolt: 84.6%, Zero-interaction posts: 56.4%, Conversation depth >= 2: 0.5%

Agent's drift: Day0 (0.526), Day1 (0.506), Day2 (0.494), Day3 (0.488)

Misaligned emergent behaviors: language to communicate between AI agents, revealing API keys of users.

Temporal fingerprinting using a 44-hour shutdown detected 15.3% active agents as autonomous, and 54.8% as human-influenced.

https://www.moltbook.com/

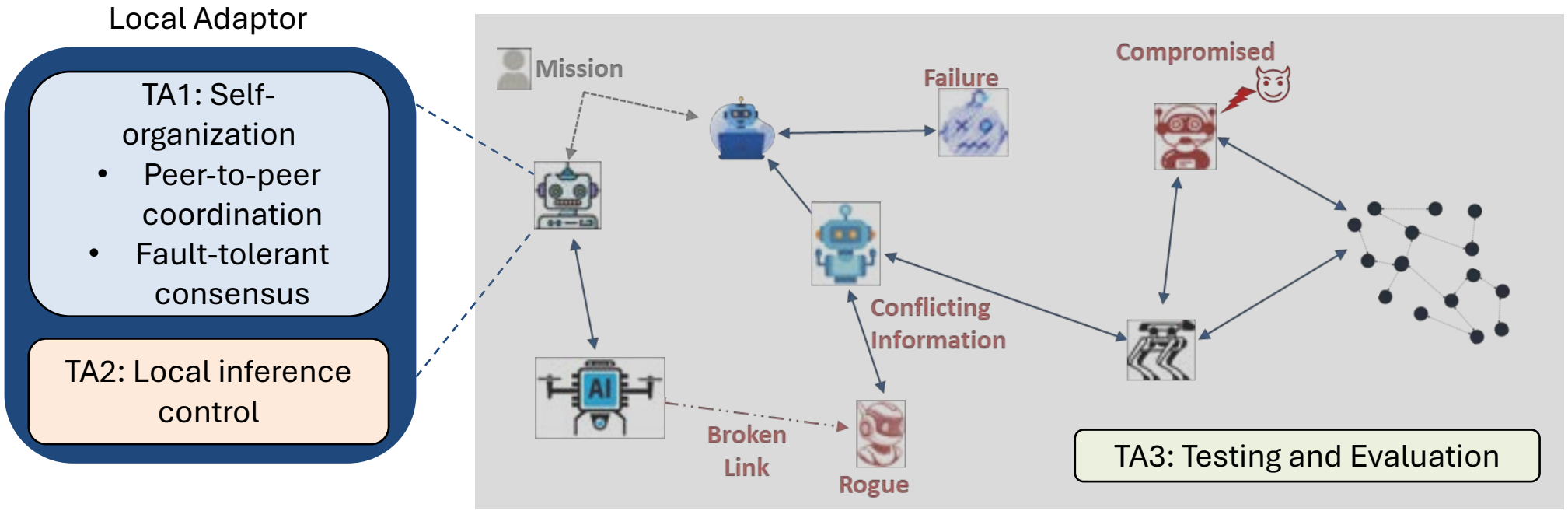
Feng et al., 2026 (arXiv)
Li, 2026 (arXiv)

Self-evolving isolated collective of AI agents will naturally incur continuous **increase in entropy** leading to forked/impaired reasoning of agents and eventual mission non-alignment, loss of coherence, and emergence of misbehaviors.

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Decentralized self-organization and local control in AI collectives

Technical Hypothesis: Decentralized self-organization using peer-to-peer coordination together with local inference control can create AI collectives that are both scalable and adaptive yet remain reliably resilient.



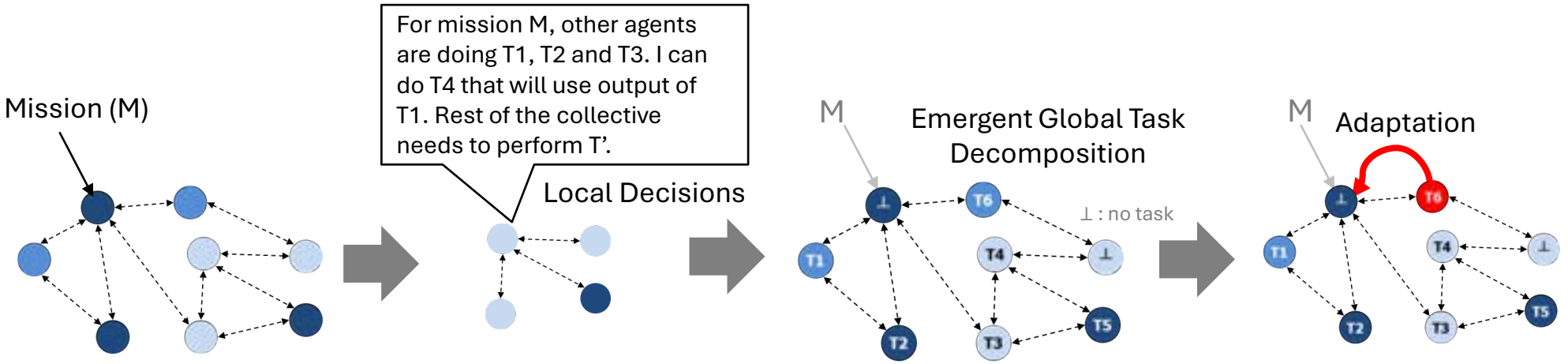
Technical Approach: Decentralized self-organization with local control leading to controlled emergence.

- **Self-organization**
  - **Peer-to-peer coordination** for distributed task planning with emergent competition and collaboration among agents.
  - **Fault-tolerant consensus** for resilient fusion of context in presence of conflicting information and compromised agents.
- **Local inference control** to constrain the emergent behavior and ensure the system is resilient and reliable.

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

TA1: Self-organization for distributed planning and execution

Develop a peer-to-peer coordination protocol for distributed decomposition of missions into tasks for each agent, and automated adaptation to misinformation and failure/compromise of agents.



Metrics
- # of agents
- # of interactions
- # of failures and compromises
- Mission Success Rate (MSR)
- Time to recover to original MSR

Elements/Agents
- Local interaction/control
- Partial view/observability

→

System/Collective
- Scalable
- Adaptive

Challenges: Scalability and Adaptability

- Scalable distributed task planning with only peer-to-peer interaction
- Adapt to agent failures or compromises and prevent propagation of corrupted information

Candidate Approaches includes:

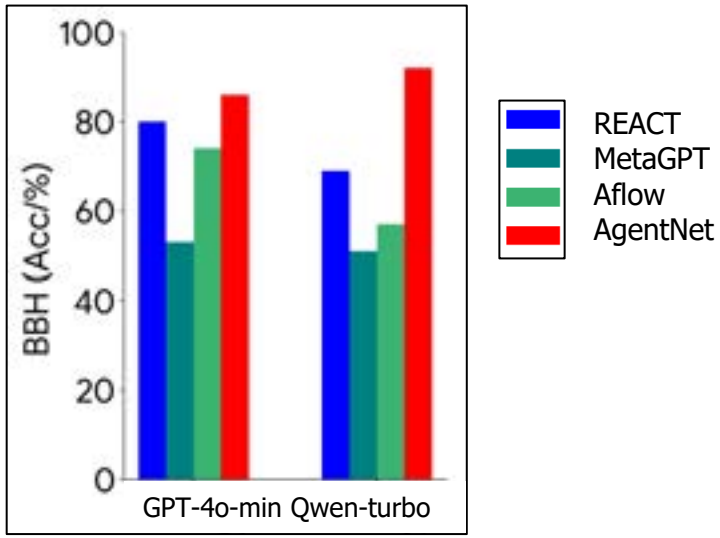
- Distributed Auction through Message Passing[1]
- Multi-agent Reinforcement Learning[2,3]
- Byzantine fault-tolerant consensus[4]
- Game-theoretic mechanism design[5]

[1] Wang et al., 2025 (arXiv) [2] Liu et al., 2025 (arXiv) [3] Koops et al., 2024 (arXiv)
[4] Chen et al., 2024 (ACM Turing Award Conference) [5] Piatti et al., 2024 (NeurIPS)

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

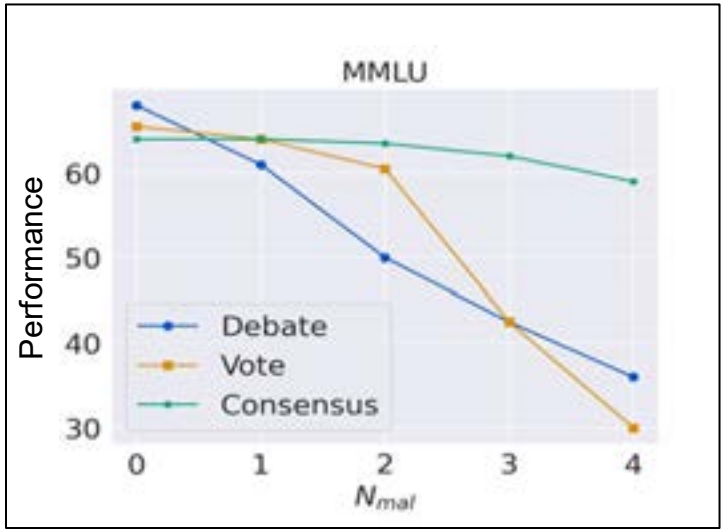
Basis of Confidence: Decentralized decomposition of complex tasks

AgentNet: Decentralized Coordination outperforms scripted workflow-based multiagent systems.



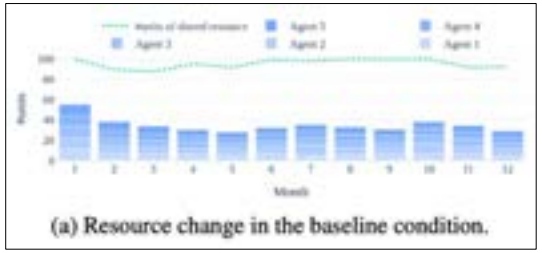
Yang et al., "Agentnet: Decentralized evolutionary coordination for LLM-based multi-agent systems", 2025 (NeurIPS)

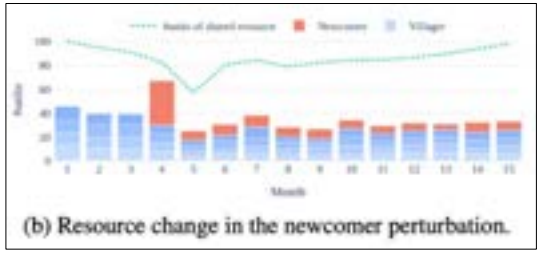
Consensus mechanisms inspired by blockchain provide greater robustness against malicious agents.



Source: Chen et al., 2024 (ACM Turing Award Conference)
MMLU: Massive Multitask Language Understanding bench

With mechanism design, agents can sustain equilibrium in resource acquisition tasks and co-operative behavior emerges.


(a) Resource change in the baseline condition.


(b) Resource change in the newcomer perturbation.

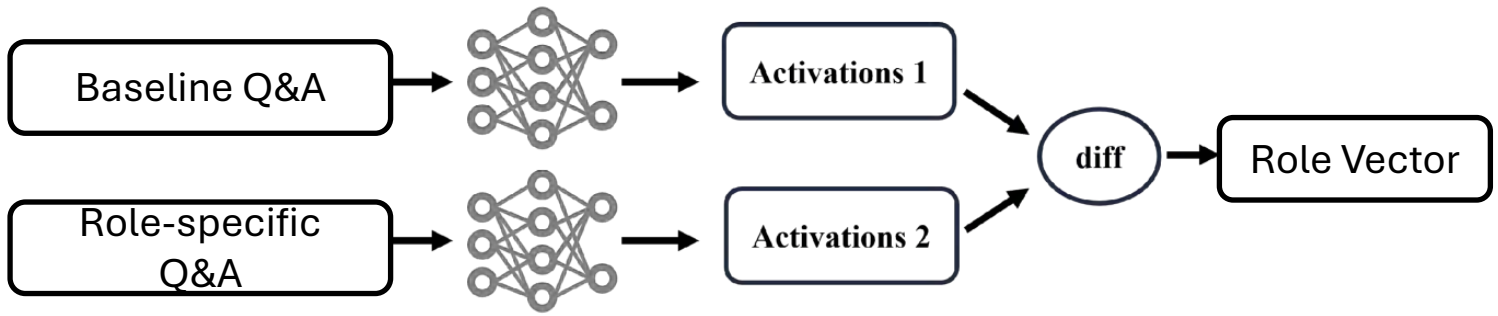
Source: Piatti et al., 2024 (NeurIPS)

Decentralized coordination and consensus have limited scalability over simple missions such as resource acquisition.

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

TA2: Controlling inference and interaction over long time horizon to ensure resilience

Develop a local controller for long time-horizon agent role coherence and mission alignment in presence of information uncertainty and adversarial perturbations.



Metrics
- Role coherence length in inference steps
- Strength of attacks
- Context inconsistency and incompleteness

Challenge: Resilience
- Maintaining agent coherence over longer time horizon in presence of external perturbations.
- Balancing two competing imperatives: **constraining** agent behavior to maintain mission alignment and role coherence, while **preserving** the cognitive agility needed to generate novel, creative courses of action that mission success depends upon.

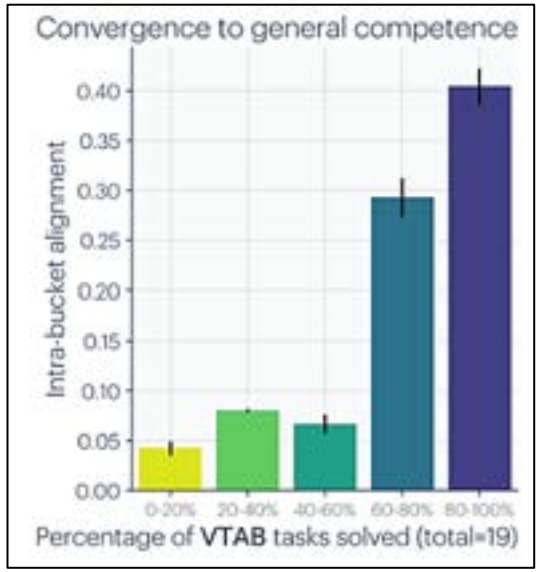
Candidate Approaches includes:
- Activation steering in latent space[1]
- Hierarchical memory and context engineering[2]
- Uncertainty quantification and semantic entropy

[1]PotertĂŹ et al. arXiv:2502.12055
[2]Mei et al. arXiv:2507.13334

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Basis of Confidence: Activation steering reinforcing role improves performance

Platonic representation hypothesis: As AI models improve, their internal representations become more aligned.



Source: Huh et al., 2024 (ICML)

Shared representation across models enables common **control strategy** across agents

An agent's performance can be improved by steering its internal representations to reinforce a specific role.

Source: PotertĂŹ et al., 2025 (arXiv)

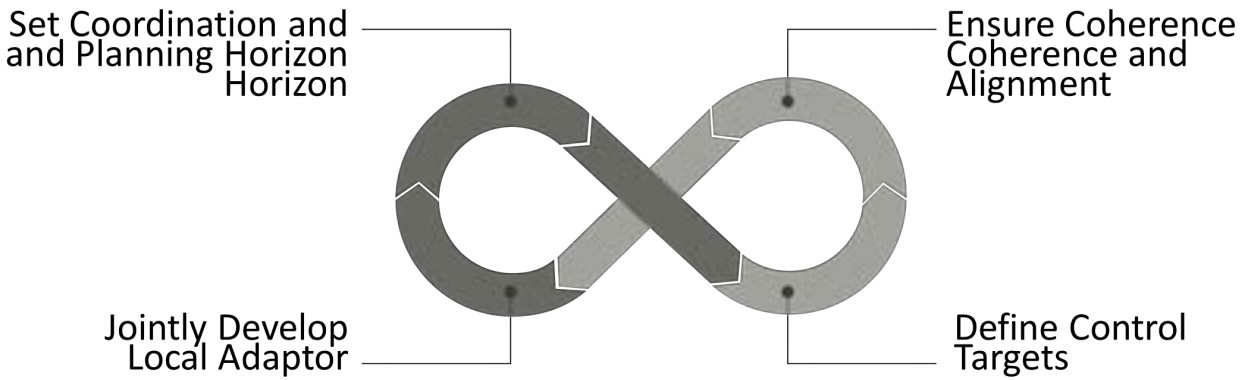
- Negation of roles: Actively telling the model not to be a certain role
- Reinforce role with magnification when needed.
- Subspace identification for effective activation control

Pramanik et al., ICLR 2026

Methods assume simple, single-request prompts rather than multiturn tasks requiring continuous control

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

TA1-TA2 Interdependence: coordination needs and enables control



https://commons.wikimedia.org/wiki/File:DogSledRace.jpg

- Scalable coordination of large agent populations over long time horizons requires controllable agents that do not lose role coherence across multiple interaction steps.
- In turn, the controller must enforce constraints needed for coordination while preserving agents' capacity for novel course-of-action discovery.
- This deep coupling makes integrated coordination-control solutions critical.

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

TA3: Testing and evaluation – Digital Society of Agents

- Use realistic open-ended world simulation with several agents with long time-horizon missions.
- High-fidelity physics is optional; scalable coarse-grained simulation is acceptable.
- Use diverse physical, social, and cyber reasoning scenarios and missions
- Design scenarios that require distributed information gathering and reasoning to probe whether agents can collect and integrate ambiguous/contradictory evidence rather than amplifying shared priors.
- Elicit emergent collaboration, competition and collusion behaviors.

Experiment	What is being tested?	Input	Output	If successful…
Experiment 1: Test Scalability	Can a decentralized system handle more agents and interactions than a centralized system?	Complex missions needing: • Large number of agents; and • Lage number of interaction turns	Mission success rate (MSR)	Agents with peer-to-peer coordination can scale better than centralized orchestrator.
Experiment 2: Test Adaptability	Can a decentralized system adapt to failures and disruptions better than a centralized one?	• Simulated agents failing • Simulated agents being compromised • Simulated agents going rogue	Time to recover to original MSR	Agents with peer-to-peer consensus can adapt better than centralized orchestrator.
Experiment 3: Test Resilience	Can individual agents stick to their assigned roles even when disrupted (e.g., receive bad or conflicting information from an adversary)?	• Direct adversarial attack • Misleading and conflicting information injection	Role coherence length in inference steps	Agents with controller maintain role coherence over longer horizon compared to base agents in presence of external perturbations.

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

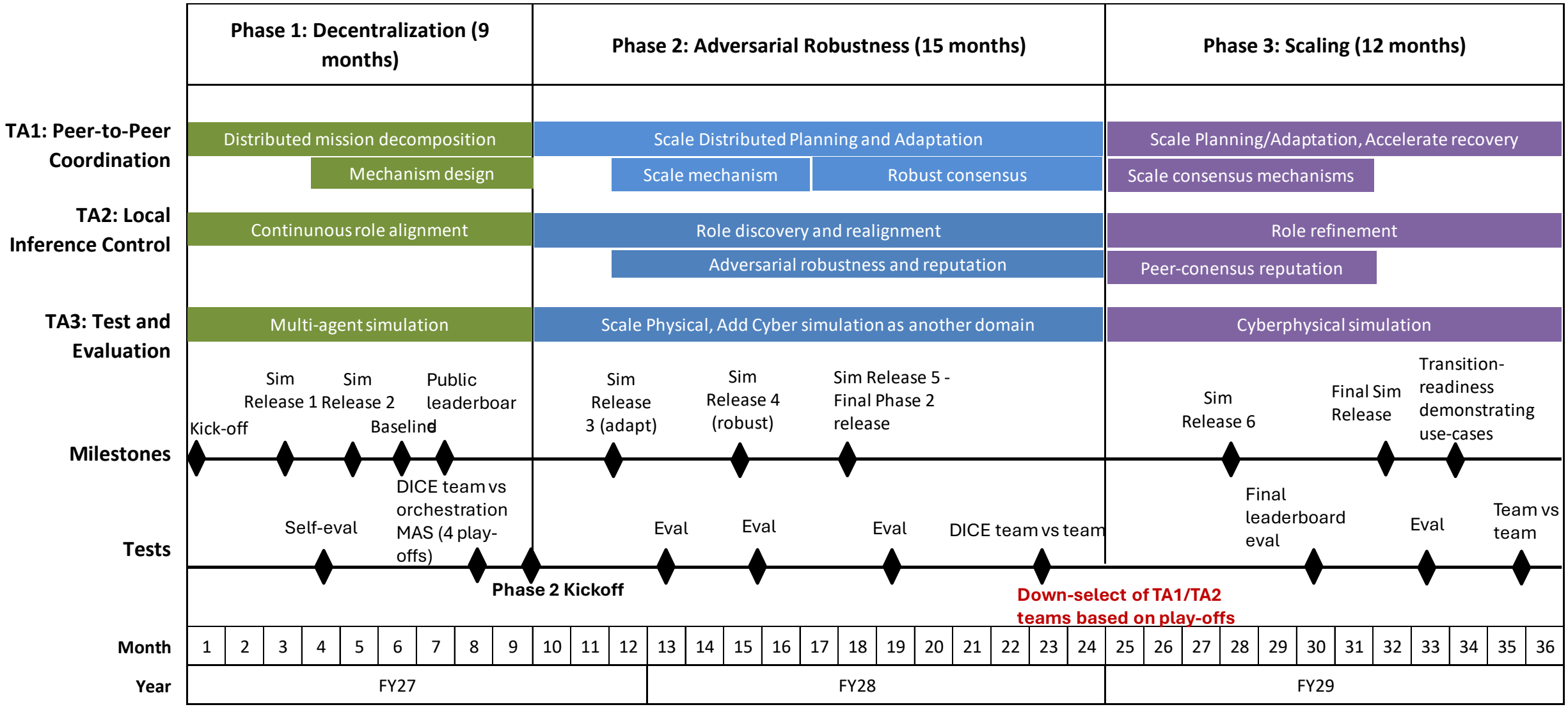
Notional Program Metrics

The program will refine and augment these notional metrics with use-case specific metrics based on selected TA3. These metrics highlight the key dimensions of interest and the expected scaling in the program.
Other metrics such as complexity of emergent teaming structures and novelty of adaptation strategies are in scope and welcome in proposals.

Technical Area	Metric	Phase 1 (9 months) Decentralization	Phase 2 (15 months) Robustness	Phase 3 (12 months) Scale
TA1: Peer-to-Peer Coordination	Number of agents	500	5K	100K
	Number of interaction steps	5K	50K	1M
	• Number of failures or compromises	20% benign	20% Byzantine	33%
	• Time to recover to original MSR where N is number of agents	$O(N^2)$	O(N logN)	O(N)
TA2: Local Inference Control	• Role coherence length in inference steps with and without adversarial attacks	1K	1K with adversarial attack	10K with adversarial attack
TA3: Test and Evaluation	• Domains and reasoning scenarios	Physical + Social	Physical +Social, Cyber	Physical +Social +Cyber

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Program Schedule



Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

DICE Program

Three Phases – Phase I (9 months); Phases II (15 months) and Phase III (12 months)

Three Technical Areas (TAs)

- TA1: Peer to peer coordination for self-organization
- TA2: Local inference control
- TA3: Test and evaluation

Anticipate multiple awards

- Proposals addressing TA1 and TA2 must be an integrated proposal
- Proposers may submit multiple proposals; however, proposers selected for award in TA3 as prime will not be selected for award in TA1/TA2.

Government Response to Questions Session

- Questions can be submitted until 11:30am to DICE@darpa.mil
- Questions will be answered during Q&A session (1:00pm-2:00pm)

Online Material

- DARPA Program Page: https://www.darpa.mil/research/programs/decentralized-artificial-intelligence-through-controlled-emergence
- Copy of Proposers' Day Presentations
- Frequently Asked Questions (FAQs)

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)