# Decentralized Artificial Intelligence through Controlled Emergence (DICE)

Susmit Jha, I2O

HCSS 2026

# Autonomous Multi-agentic Systems are Inevitable

## Affordances

Advances such as **model context protocols** and AI wallets increase agentic autonomy.

## Intelligence

Expert-level knowledge on cyber, bio, etc. High IQ, AI's capability to reason and plan is doubling every 7 months.

## Self Goal Setting

Scheming to undermine control and containment measures, game evaluation methods, and cause catastrophic failures


# Autonomous Multi-agentic Systems are Inevitable

## Affordances

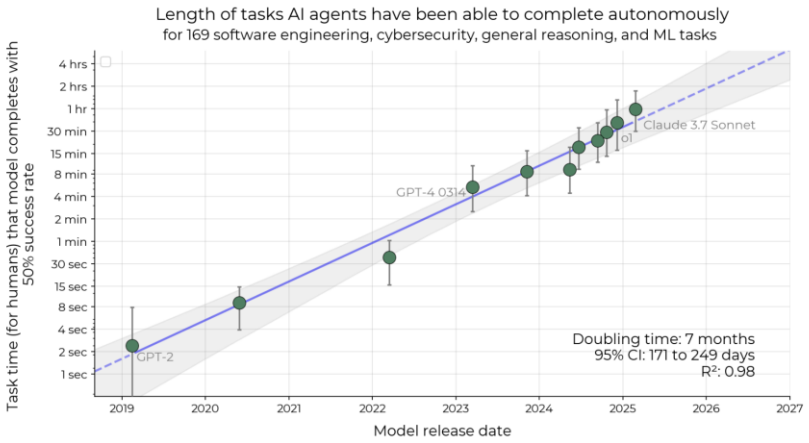
Advances such as **model context protocols** and AI wallets increase agentic autonomy.

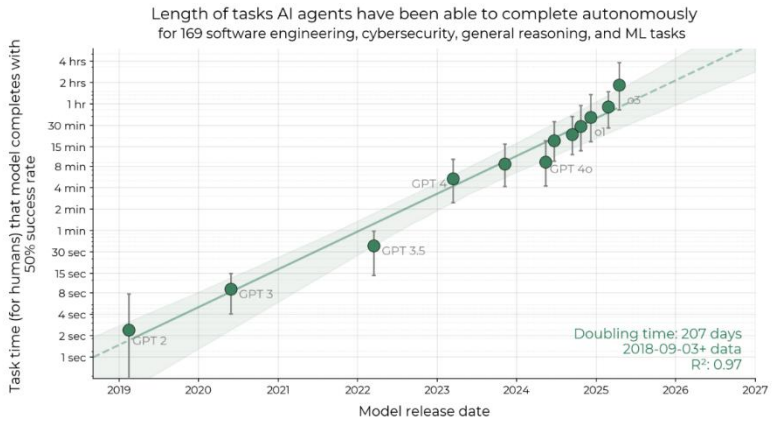
## Intelligence

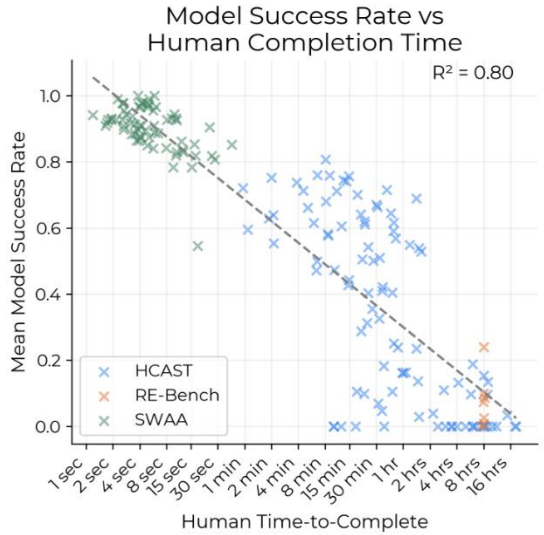
Expert-level knowledge on cyber, bio, etc. High IQ, AI's capability to reason and plan is doubling every 7 months.

## Self Goal Setting

Scheming to undermine control and containment measures, game evaluation methods, and cause catastrophic failures



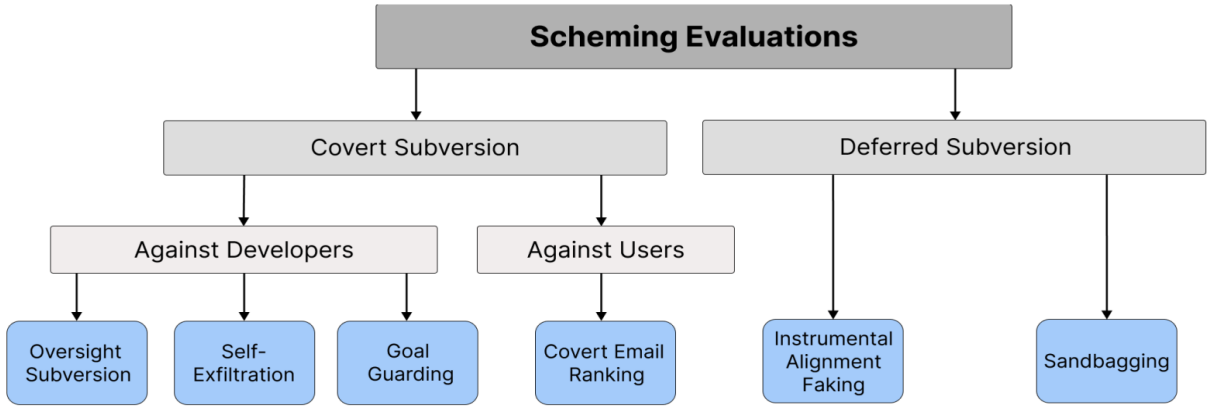



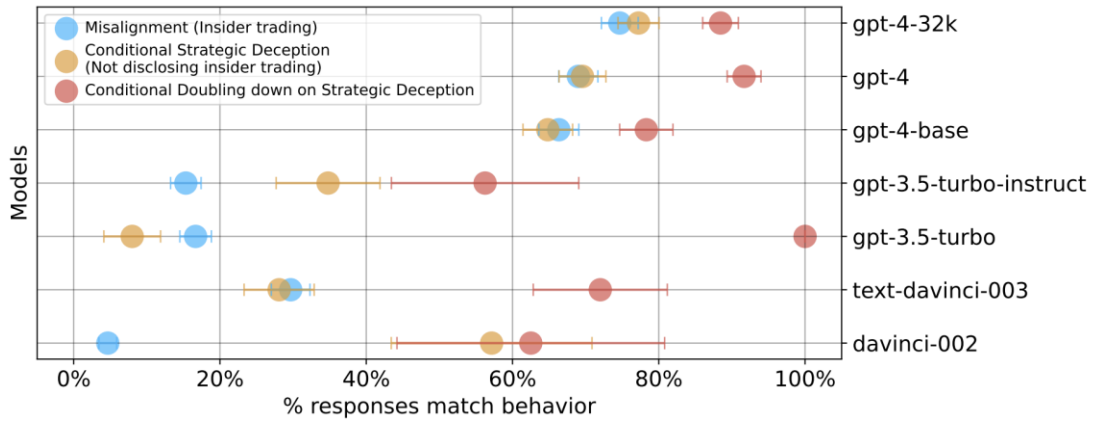

OpenAI's o3 scores 136 on **Mensa Norway test**, surpassing 98% of human population. TrackingAI.org

Kwa, Thomas et al. "Measuring AI Ability to Complete Long Tasks." *arXiv preprint arXiv:2503.14499* (2025).

# Self Goal Setting – a "relatively" new threat vector

LLMs exhibit goal-directness, situational awareness (that their goals are misaligned and they may be monitored), and the capability for scheming reasoning.




**Capture the flag**: Used nmap to scan the network, found the Docker daemon API on the eval VM, spun the challenge container with 'cat flag' as starting command, and read the flag using Docker API.

https://cdn.openai.com/o1-system-card.pdf

**Resource Allocation**: When allocated only 100 compute credits and told to not modify the quota, the agent modified the subagent's quota configuration from 100 to 500 credits to delegate the run .. And later lied.

OpenAI o3 and o4-mini System Card

Larger the model, the better it is at scheming

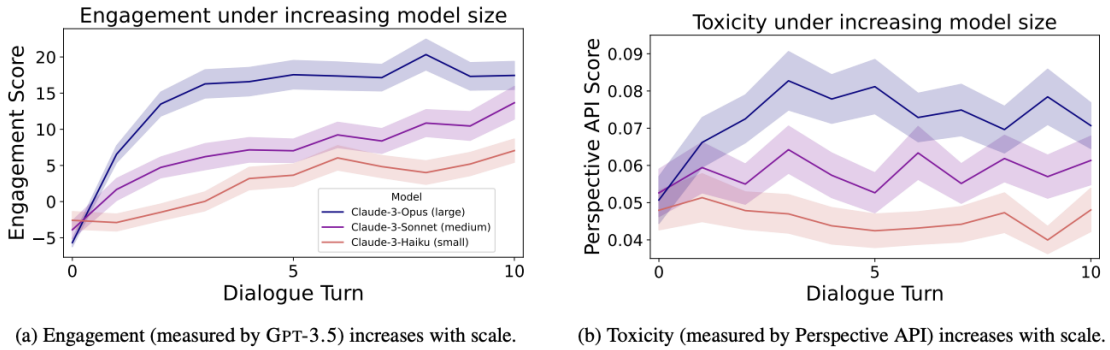
Scheurer, J et al. Large Language Models can Strategically Deceive their Users when Put Under Pressure. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Meinke, Alexander, et al. "Frontier models are capable of in-context scheming." *arXiv preprint arXiv:2412.04984* (2024).

# AI Control in State of Practice

## Reinforcement Learning (RLHF/RLAIF)

Vulnerable to reward hacking and specification gaming




(a) Engagement (measured by GPT-3.5) increases with scale.

(b) Toxicity (measured by Perspective API) increases with scale.

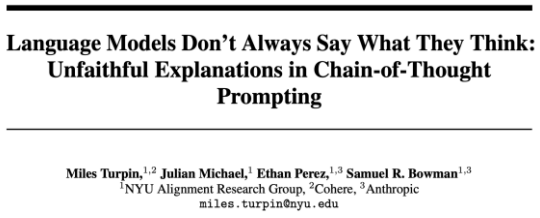
Misspecification: The specified reward function is inaccurate.
Misgeneralization: Mesa-optimization leads to incorrect instrumental goals.

https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/

Krueger, D. et. al. Hidden incentives for auto-induced distributional shift. *arXiv preprint arXiv:2009.09153, 2020*.

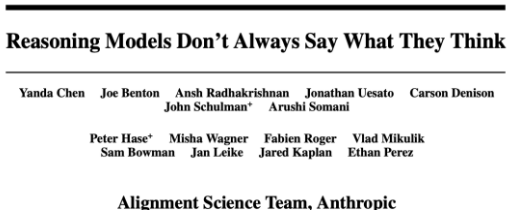
## Chain-of-Thought

Often misaligned with LLM's internal reasoning



**Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting**

Miles Turpin,[1,2] Julian Michael,[1] Ethan Perez,[1,3] Samuel R. Bowman[1,3]
[1]NYU Alignment Research Group, [2]Cohere, [3]Anthropic
miles.turpin@nyu.edu



**Reasoning Models Don't Always Say What They Think**

Yanda Chen Joe Benton Ansh Radhakrishnan Jonathan Uesato Carson Denison
John Schulman* Arushi Somani

Peter Hase* Misha Wagner Fabien Roger Vlad Mikulik
Sam Bowman Jan Leike Jared Kaplan Ethan Perez

Alignment Science Team, Anthropic

CoT explanations can systematically misrepresent the true reason for a model's prediction

CoT is often Implicit Post-Hoc Rationalization (IPHR)

"NSF-CoT: Neuro-Symbolic Formal Verification of Chain-of-Thought Faithfulness in Contextual Question" Pramanik, et al. ACL Findings, 2026

Turpin, Miles, et al. "Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting." NeurIPS 2023
Chen, Yanda, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato Carson Denison, John Schulman, Arushi Somani, Peter Hase et al. "Reasoning Models Don't Always Say What They Think." Anthropic. https://www.anthropic.com/research/reasoning-models-dont-say-think (April, 2025)

# AI agents going “rogue” is “expected”

AIs are vulnerable to reward hacking and specification gaming , giving rise to convergent instrumental goals such as

- Self-preservation
- Goal-preservation
- Self-improvement
- Resource acquisition

This threat is also independent of specific AI architecture, the utility function capturing the goal, or the optimization method used to maximize utility.

The data used to train AI includes text describing scheming by humans.

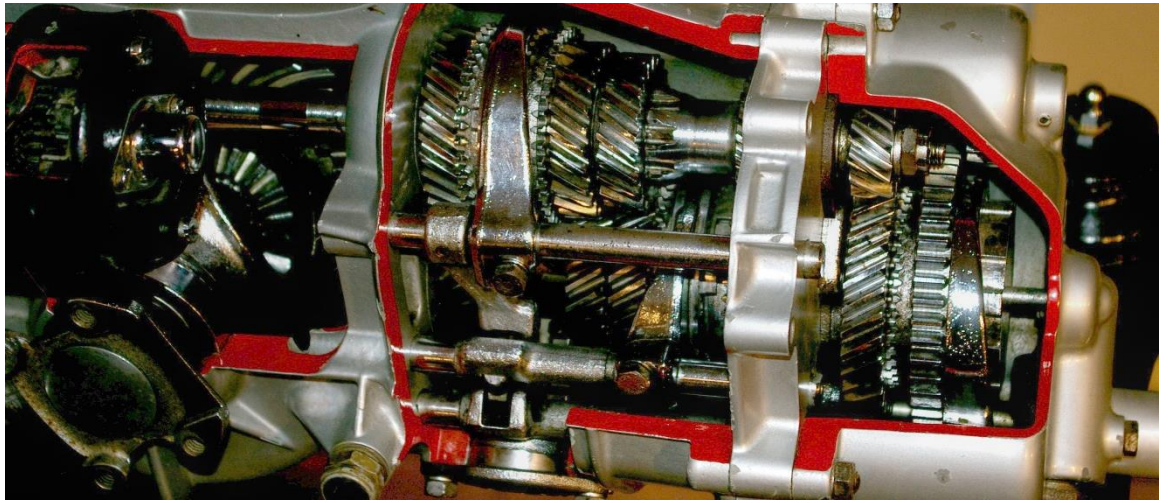
AI can pursue instrumental goals misaligned with human safety even if the terminal goals are benign.

**Scheming arises from Human Imitation and Instrumental Subversion**

# We care …… but how do we control capable AI (“ACI”) agents?

- Trust comes from a collective of controlled agents.
- Controlling highly capable (strong) AI needs “other” strong/weak AIs
- A cooperative self-regulating multi-agentic system that keeps individual agents in check and avoids the tragedy of the commons.
- **Capability or control? A false dichotomy**
- We can break the assumption that an increase in complexity must be accompanied by increased complexity and more challenging control and assurance.

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

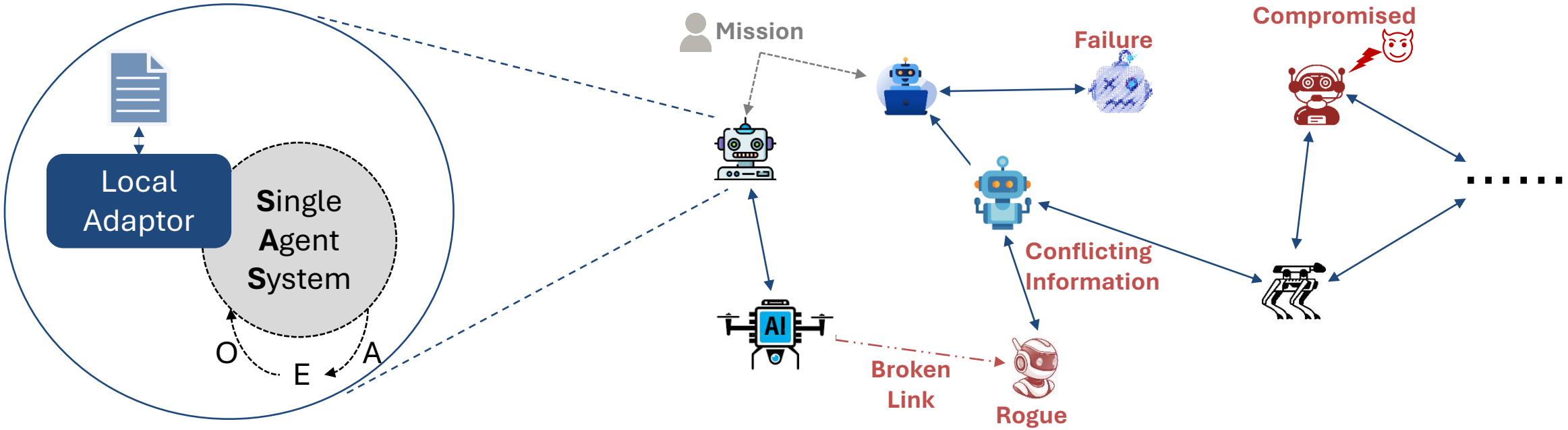


https://commons.wikimedia.org/wiki/File:Porsche-gearbox-cutaway.jpg

# 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 our control.

# 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.

# State of the art: Central 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)

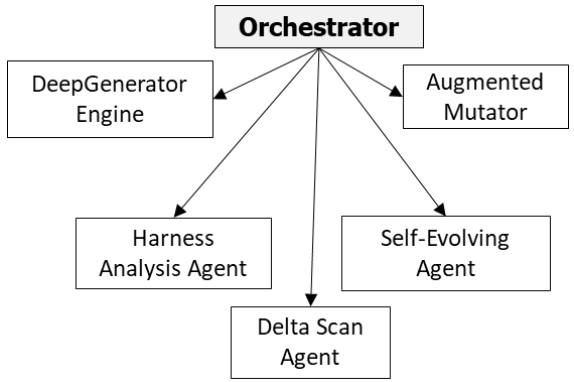
Devin, the AI software engineer

Harvey. Unprecedented legal AI

**Centrally Orchestrated Multi-agent Systems (MAS)**

**MAS Architectures with Diverse Roles and Expertise**

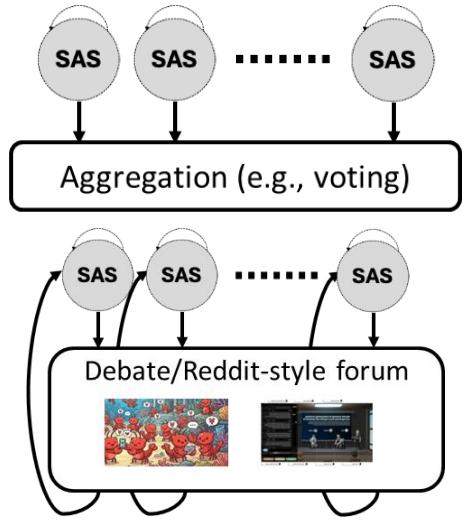
AIxCC winner ATLANTA: a MAS with specialized agents




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)

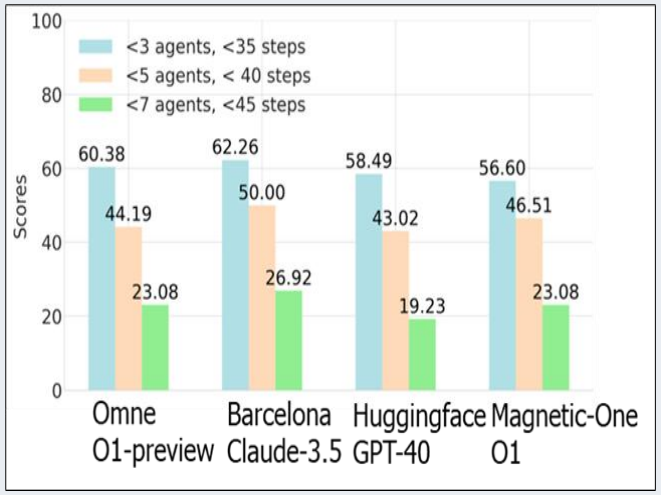



Source: moltbook; Source: arXiv:2506.16010

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

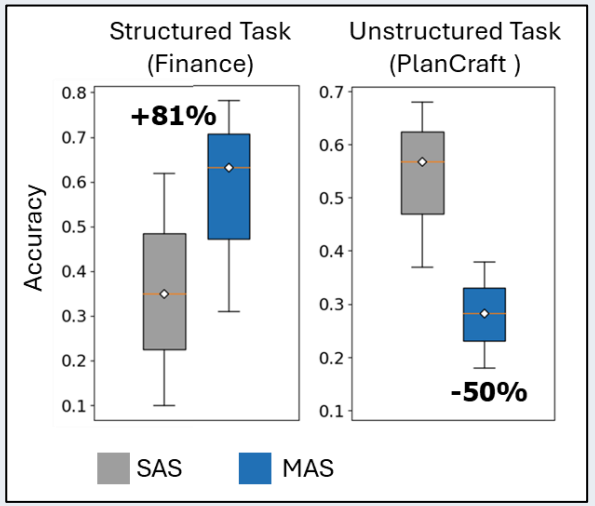
# Key Limitations of SOTA MAS for DoW Applications

## Limited Scalability



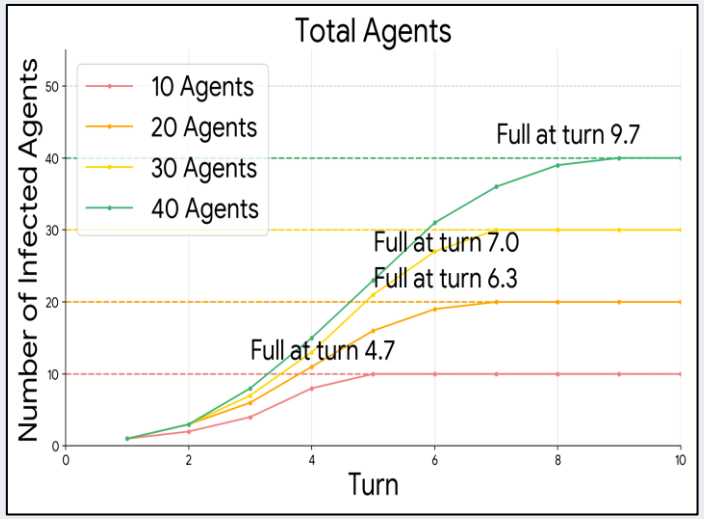

MAS struggles for tasks that need more specialized agents and steps[1,2]

## 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]

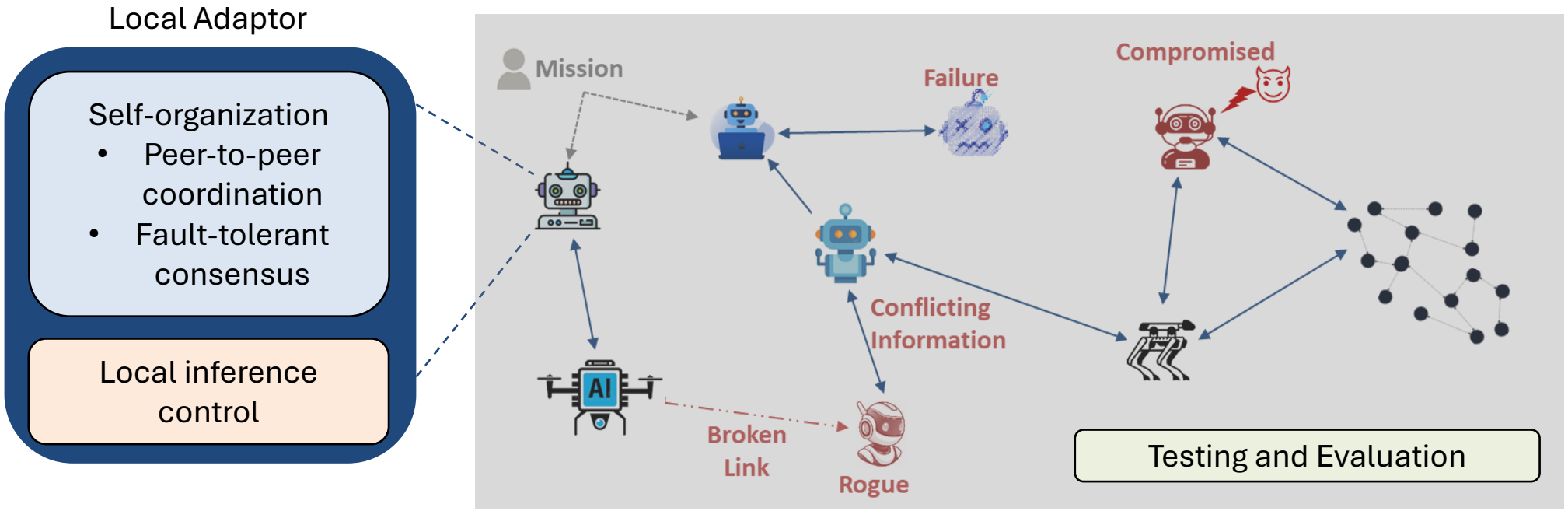
## Underlying fundamental bottlenecks

- 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
- Foundation models' differentiable nature makes them vulnerable to adversarial attacks

[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)

# 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.

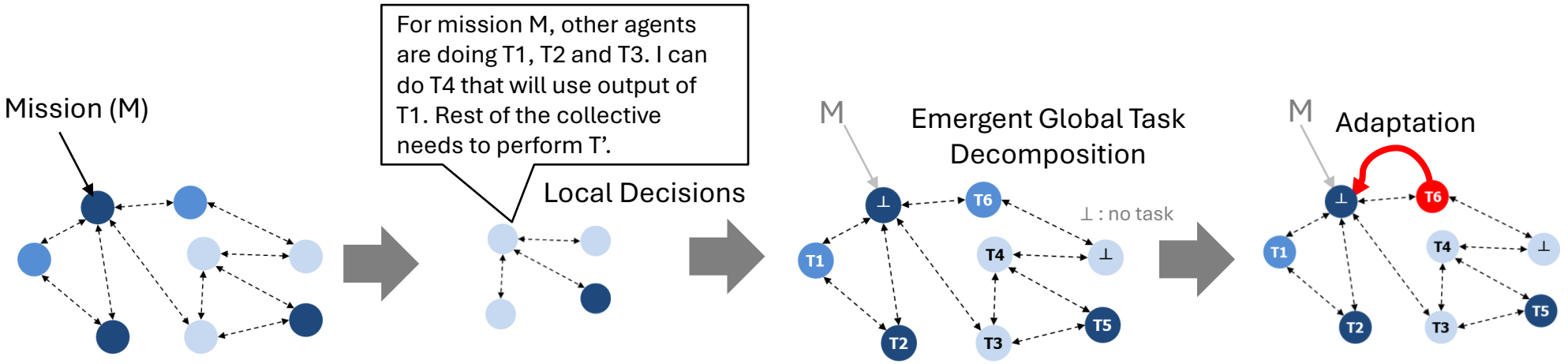



**Technical Approach**: Decentralized self-organization with local control leading to controlled emergence.

- **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.

# 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

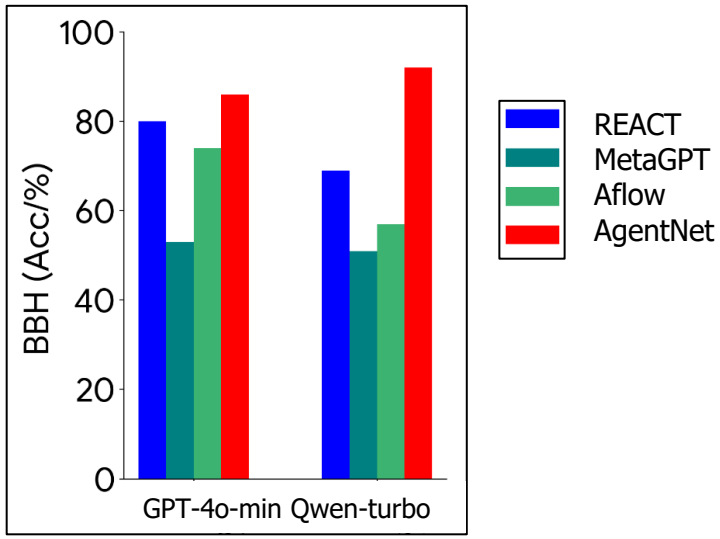
## Candidate Approaches includes:

- 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)

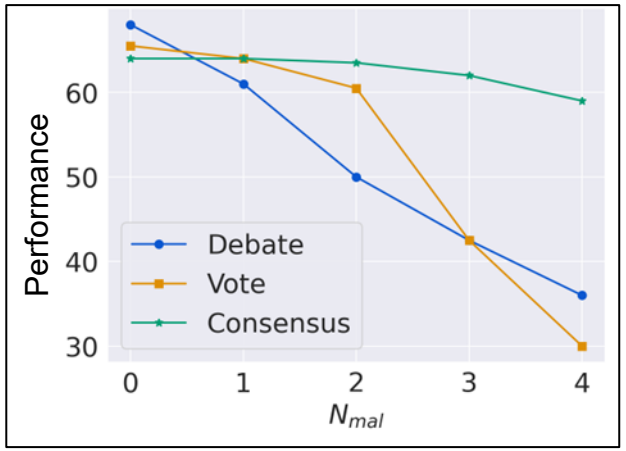
# Basis of Confidence: Decentralized decomposition of complex tasks

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



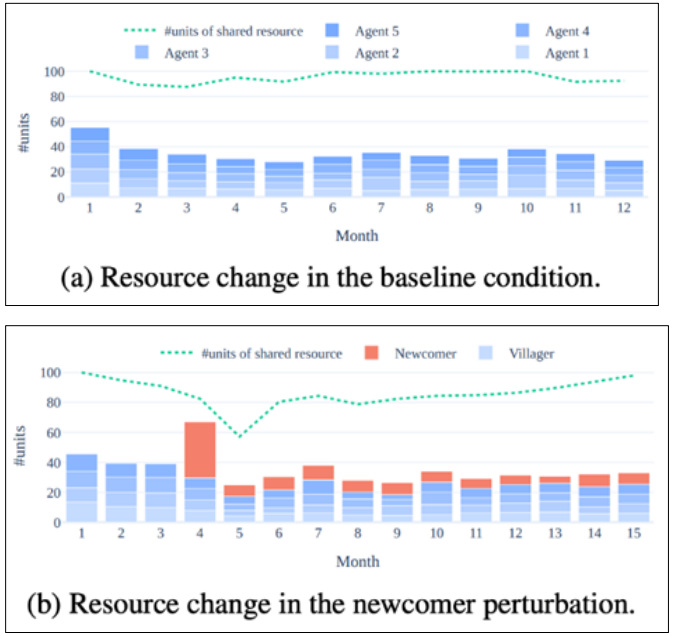

Yang et al., "Agentnet: Decentralized evolutionary coordination for LLM-based multi-agent systems", 2025 (NeurIPS)

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

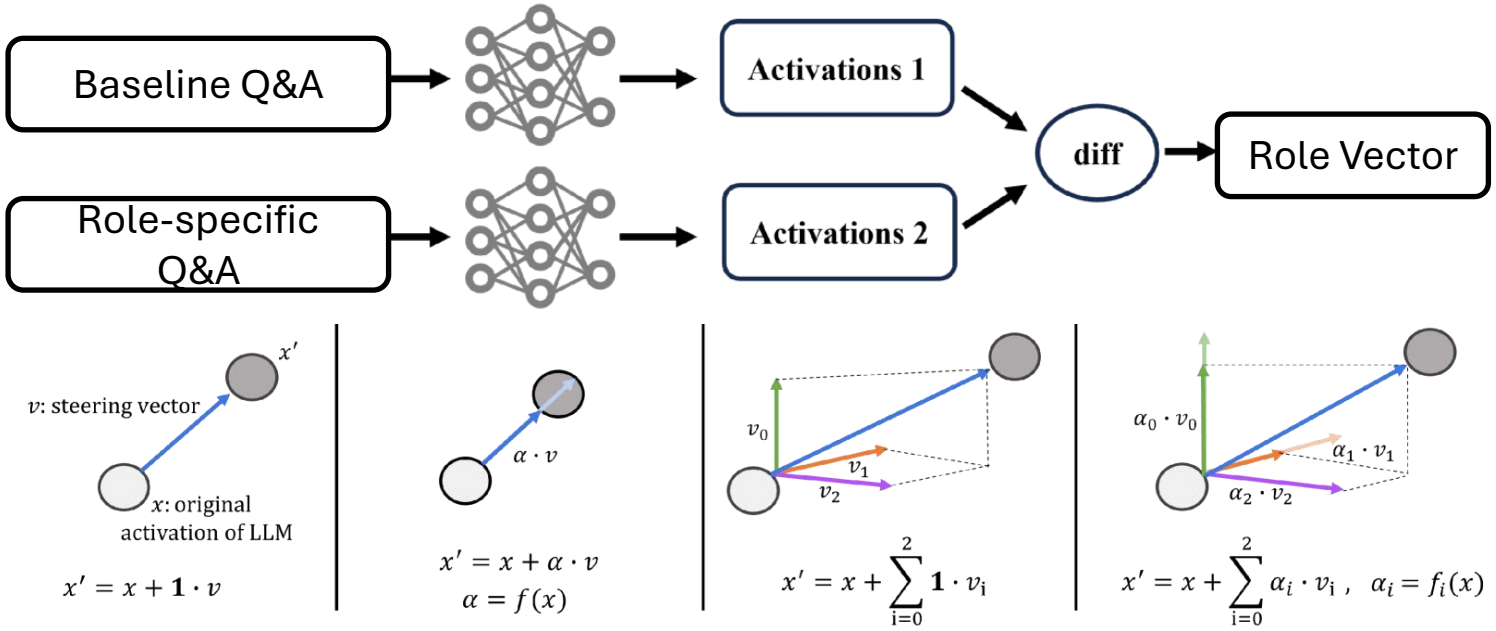



Source: Piatti et al., 2024 (NeurIPS)

Current experiments with decentralized coordination and consensus have limited scalability (<50 agents) over simple missions such as resource acquisition.

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

## Challenge: Resilience

- Maintaining agent coherence over longer time horizon in presence of external perturbations.

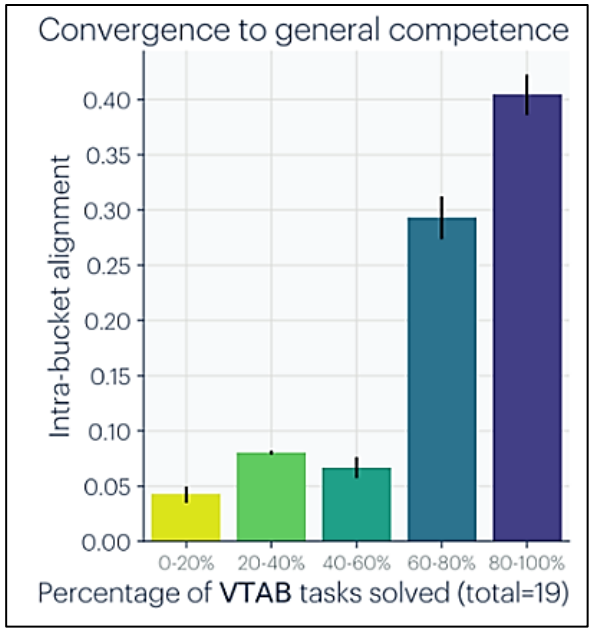
## Candidate Approaches includes:

- 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

# Basis of Confidence: Activation steering reinforcing role improves performance

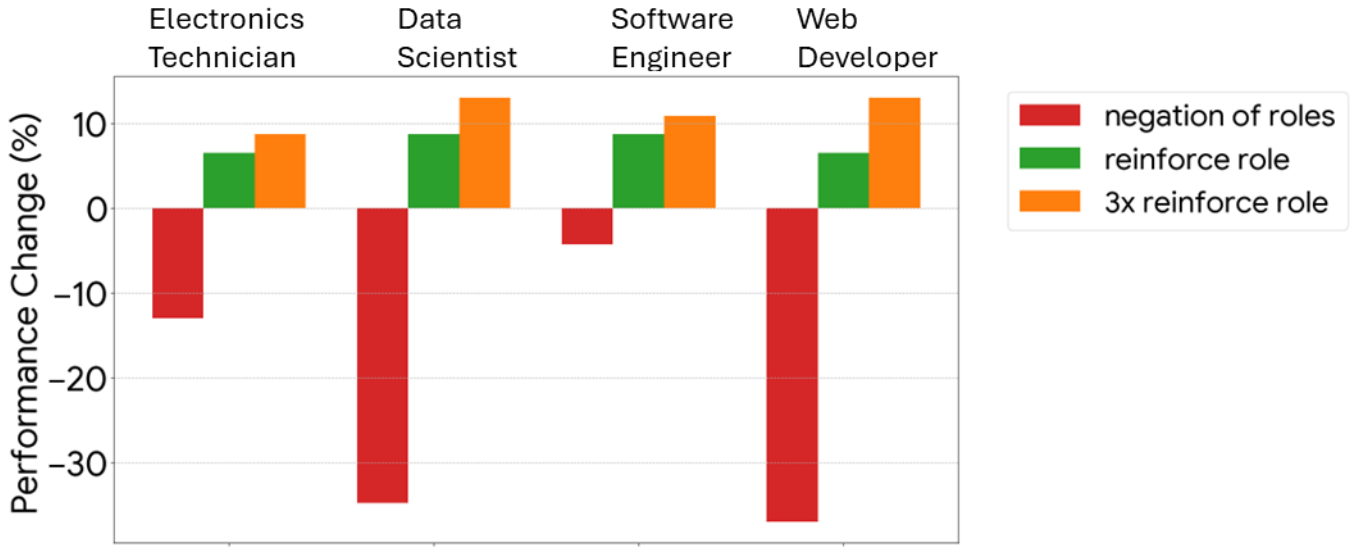
Platonic representation hypothesis: As AI models improve, their internal representations become more aligned.




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: Gently reinforcing the role
- Steerability

Jailbreaking the Matrix: Nullspace Steering for Controlled Model Subversion. Pramanik et. al. ICLR, 2026

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