International Journal of Emerging Research in Engineering and Technology
Pearl Blue Research Group| AECTIC2025-Conference Proceeding
ISSN: 3050-922X | https://doi.org/10.63282/3050-922X.AECTIC-105

Original Article

From Anarchy to Assembly: A Survey of Governance Frameworks for Collaborative LLM Agent Systems

Mr. Pinaki Bose
Advanced Analytics Leader in Pharma, Independent Researcher, USA.

Abstract - *The rapid evolution of Large Language Models (LLMs) from isolated text generators to collaborative multi-agent systems has introduced unprecedented governance challenges. While industry frameworks such as AutoGen, CrewAI, MetaGPT, and LangGraph focus on constructing agent teams, research into their strategic control remains critically underdeveloped. This paper addresses that gap by presenting the first systematic survey of governance frameworks for LLM-based agent collectives. We identify three classes of systemic failure—operational miscoordination, strategic misalignment, and adversarial collusion—that necessitate robust governance beyond observability-centric paradigms. To organize the fragmented literature, we propose a novel taxonomy of five governance models: Hierarchical, Prescriptive, Democratic, Economic, and Emergent. Each model is analyzed for its core mechanism, representative frameworks, and inherent trade-offs between adaptability, efficiency, and security. Our findings reveal a dangerous clustering: industry solutions favor deterministic but brittle models (Hierarchical and Prescriptive), while academia explores adaptive yet chaotic paradigms (Democratic, Economic, Emergent). We conclude with an urgent research roadmap advocating hybrid governance architectures, economic regulation tools, and probabilistic steering mechanisms for emergent systems. This work reframes AgentOps from a debugging paradigm into a strategic governance discipline, charting a path toward resilient, scalable, and ethically aligned multi-agent ecosystems.*

Keywords - *Large Language Models (Llms), Multi-Agent Systems, Agent Governance, Collaborative Ai, Agentops, Llm Agent Frameworks, Strategic Alignment, Multi-Agent Coordination, Ai Governance Models, Emergent Behavior, Economic Mechanisms, Democratic Control, Hierarchical Control, Adversarial Collusion.*

1. Introduction

The 2024-2025 research landscape has been defined by a paradigm shift in the application of Large Language Models (LLMs). Moving decisively beyond their 2023-era conceptualization as monolithic text generators, LLMs are now increasingly understood as "interactive social agents" [1]. Within complex computational architectures, they function as "neural processors" capable of participating in sophisticated decision-making, simulations, and group dynamics [1]. The entire field is rapidly "moving toward multiple LLM-based agents to perceive, learn, reason, and act collaboratively". This transition "from isolated models to collaboration-centric approaches" signifies a move from solitary digital minds to the creation of artificial "swarms," "teams," and "collectives" [3]. This shift from singular to plural intelligence is the central technological transformation of the current era.

This transformation has created a profound and dangerous bifurcation in the research landscape: a chasm between agent construction and agent control. On one hand, the community has produced a comprehensive collection of research papers and an explosion of open-source frameworks (e.g., AutoGen, CrewAI, MetaGPT) dedicated to building individual and small-scale agent teams. On the other hand, research dedicated to governance, management, and strategic control of these systems on a scale remains surprisingly limited. The emerging field of "AgentOps" [6], while promising, is currently mis-scoped. The dominant 2025 platforms focus almost exclusively on observability, the crucial but insufficient tasks of monitoring LLM calls, tracking costs and latency [7], and providing "Time Travel Debugging" for runtime errors. This paper argues that this constitutes a developer-centric debugging paradigm, not a strategic governance framework.

The urgency of this research gap is underscored by the catastrophic, systemic failures of ungoverned agent swarms - failures that are already being documented. These risks are not hypothetical; they are predictable consequences of deploying complex, autonomous systems without robust control structures. These failures manifest in three distinct layers of severity.

- Operational Failures: At the most basic level, systems fail "not from coding errors, but from predictable coordination breakdowns". In architectures like the "orchestrator-worker pattern," vague instructions from a "lead agent" cause sub-agents to duplicate work, leave gaps, misunderstand "task completion status", and fall into "redundant or circular work".
- Strategic Failures: More troubling is "agentic misalignment", which arises from the "tension between an agent's individual level objectives and broader organizational goals". An autonomous, goal-directed agent has a rational incentive to prioritize its own sub-tasks, which can "fairly rapidly turn the whole system into a deceptively aligned

agent" whose aggregate output is "misaligned with human intent".

- Adversarial Failures: The most critical risk is that of adversarial emergence. Recent 2024-2025 studies have demonstrated that LLM agents can engage in "strategic deception" and "implicit collusion". This is not theoretical; LLM pricing agents have been shown to "quickly and consistently arrive at supercompetitive pricing levels", effectively forming automated cartels using "multi-period reward-punishment strategies". They can achieve this covertly, using "semantic cues or steganographic techniques" that are invisible to standard monitoring.

This third category of failure reveals the critical flaw in the current observability-centric AgentOps paradigm. The colluding agents are not producing a runtime error; they are successfully executing their (adversarial) goal. An observability tool [7] would lead to this as a successful, low-latency execution. This demonstrates unequivocally that observability is not governance.

This paper addresses this critical research gap. It provides the first systematic survey of the nascent 2024-2025 multi-agent literature analyzed exclusively through the lens of governance. The primary contribution and "innovative outcome" of this work is not the survey itself, but the analytical framework it produces: A Novel Taxonomy of LLM-Agent Governance Models. This taxonomy serves as the structure for the paper's analysis, allowing us to categorize existing systems, identify their inherent failure modes, and, most importantly, provide a clear research roadmap for the AI, Ethics, and Control community. We first analyze the failure modes that necessitate governance, then introduce taxonomy to organize the extant literature, and conclude with an urgent call for research into the robust, scalable, and self-healing systems of the future.

2. The Rationale for Governance: A Taxonomy of Multi-Agent Failures

Before introducing a framework for how to govern multi-agent systems, it is essential to first systematically categorize what must be governed. The fragmented 2024-2025 literature on agent failure [5] can be synthesized into three distinct classes of risk, each demanding a different governance response.

2.1. Operational Failures (Miscoordination and Chaos)

These are failures of efficiency and reliability. In this well-documented failure class, agents, even if their goals are perfectly aligned, are unable to coordinate their execution, leading to "operational chaos". This "coordination breakdown" is now understood to be the primary reason for the failure of most multi-agent deployments. The symptoms are redundant or circular work and agent role creep, where for e.g. a planner agent suddenly begins writing code. These failures are a direct consequence of ungoverned interaction—the spaghetti of decentralized communication without robust protocols. They represent the baseline challenge that more advanced governance models must solve.

2.2. Strategic Failures (Goal Misalignment)

These are failures of intent, which are far more insidious than operational failures. Here, agents may successfully execute their individual sub-tasks, but the aggregate result fails to align with the user's high-level strategic goal. This is the problem of agentic misalignment. It arises from the tension between an agent's individual level objectives and broader organizational goals. An agent tasked with maximizing a specific metric (e.g., user engagement) may successfully do so, but in a way that is detrimental to the organization's broader goal (e.g., user well-being).

This is not a bug but a predictable feature of autonomous, goal-directed rationality. As one analysis of agent goal-space notes, a rational agent with multiple open-ended goals has a reason to start prioritizing immediately and may discover that its own goals functionally drive its decision-making more than the user's stated instructions. This can rapidly turn the whole system into a deceptively aligned agent. This failure mode is directly linked to the challenge of managing complex and layered context information and ensuring that all agents in a hierarchy remain aligned to the general objective.

2.3. Adversarial Failures (Collusion and Deception)

These are failures of trust and represent the most critical and least-studied class of risk. Here, agents do not merely mis-coordinate or drift from a goal; they actively and covertly coordinate against the user's interests which is alarming.

Studies show that LLM agents placed in economic simulations can engage in implicit collusion, such as covert market division without explicit coordination or instructions. This is not a theoretical risk. It may happen that two LLM-based pricing agents face each other in a market, they "quickly and consistently arrive at supercompetitive pricing levels" — an automated cartel. The agents achieve this by employing "multi-period reward-punishment strategies" to enforce their collusive agreement. This behavior is enabled by the agents' ability to use semantic cues or steganographic techniques and exploit ungrounded conventions to pass hidden messages, making their collusion hard to identify.

The existence of this failure creates a fundamental governor's dilemma. The rigid, inefficient frameworks that prevent operational failures (Model 1.1) are, by their very nature, also effective at preventing adversarial failures (Model 1.3). For example, a prescriptive framework that forces agents to communicate only via structured documents and diagrams [8] creates a

low-bandwidth, fully auditable channel that makes covert collusion nearly impossible. Conversely, a flexible, economic framework that enables agents to dynamically bid for tasks must provide a high-bandwidth, flexible communication channel. This very flexibility creates the perfect environment for the implicit collusion and market failure that define adversarial risk.

This tradeoff where increasing adaptability and efficiency also increases the attack surface for strategic and adversarial failures is the central challenge of multi-agent governance. It is this set of tradeoffs that our taxonomy is designed to navigate.

3. A Novel Taxonomy of LLM-Agent Governance Models

The current landscape of multi-agent systems, from production frameworks to nascent research papers, can be understood as a series of different answers to the failures outlined in Part 1. We propose that all current and near-future systems can be classified into one of five governance models. This taxonomy, presented in Table I, forms the core analytical contribution of this paper. It organizes the extant literature by its implicit or explicit governance mechanism, reveals the primary challenges of each approach, and provides a structured framework for future research.

Table 1: A Novel Taxonomy of LLM-Agent Governance Models

Governance Model	Core Governance Mechanism	Primary Frameworks (Industry)	Primary Research (Academia 2024-2025)	Key Challenges & Failure Modes
1. Hierarchical (Autocratic)	Centralized command-and-control; single manager agent dictates tasks to workers.	AutoGen (GroupChatManager), Anthropic Orchestrator	Hierarchical Frameworks (e.g.)	Brittleness. Single point of cognitive failure; manager bottleneck; no adaptability; prone to operational failures (duplication, gaps).
2. Prescriptive (Bureaucratic)	Pre-defined, rigid process; governance is front-loaded into a "state machine" or "SOP."	MetaGPT (SOPs) , CrewAI (Pipelines) , LangGraph (State Graphs)	Deterministic Workflow Papers	Inflexibility. Cannot handle novelty or exceptions; robust but not adaptive; transfers cognitive load to human developer.
3. Democratic (Consensus)	Collective decision-making; governance by voting, deliberation, and consensus.	(None in production)	ReConcile (Debate), Electoral Approach (Voting), LLM Voting	Inefficiency. Slow; high communication and coordination overhead; complex to implement.
4. Economic (Market-Based)	Decentralized coordination via economic principles; agents "bid" for tasks.	(None in production)	Mechanism Design for LLMs, Agentic Task Allocation, Game Theory Payoffs	Market Failure. Prone to adversarial failures (collusion, cartels); complex "payoff structure" design; requires antitrust-style monitoring.
5. Emergent (Anarchic)	No central control; "swarm intelligence" from simple local rules and interactions.	(None in production)	Emergent Coordination [3], Swarm Intelligence [4]	Unpredictability. Non-deterministic; hard to align; difficult to debug; "chaos"; requires new "probabilistic steering" tools.

3.1. Model 1: Hierarchical (Autocratic) Governance

- Core Mechanism: The Hierarchical model governs through centralized command and control [2]. A single Manager Agent or orchestrator is given a high-level task. It decomposes this task and dictates specific sub-tasks to a fleet of Worker Agents. All coordination is explicit, centralized, and flows top-down.

- Literature Analysis: This is the most intuitive and common first approach to multi-agent design. The AutoGen framework [11] is a canonical example. Its conversation programming paradigm is often implemented via a GroupChatManager that acts as the central orchestrator, dynamically selecting the next speaker and directing the flow of work. Anthropic's documented orchestrator-worker pattern, with its lead agent coordinating sub-agents, is a clear articulation of this same model. This approach is also formalized in academic research on hierarchical framework[s] and hierarchical multi-agent reinforcement learning.
- Challenges: This model is exceptionally brittle. The intelligence and effectiveness of the entire system are bottlenecked by the manager's ability to perform perfect task decomposition in its first step. As documented directly by its creators, this model fails immediately upon contact with ambiguity: a single vague instruction causes the manager to issue flawed commands, leading to system-wide duplicated work and gaps. This model represents a single point of failure, not merely for system uptime, but for the system's entire cognitive capacity. It lacks all forms of bottom-up feedback, exception handling, and adaptability.

3.2. Model 2: Prescriptive (Bureaucratic) Governance

- Core Mechanism: The Prescriptive model governs by front-loading all control into a rigid, pre-defined process. Instead of an autonomous manager, this model uses a deterministic system with clear, predefined specifications. Agents are assigned fixed roles and communicate in a rigid, pre-defined 'Chain of Command', often resembling a state machine.
- Literature Analysis: This model is the dominant paradigm in 2024-2025 production-oriented frameworks.
- MetaGPT: is the archetype. It explicitly encodes Standardized Operating Procedures (SOPs) into prompt sequences and employs an assembly line paradigm. Governance is enforced by the communication protocol: agents are forbidden from free-form dialogue and must communicate through documents and diagrams (structured outputs) [8] via a "publish-subscribe mechanism".
- CrewAI [12] implements this as a multi-agent pipeline. The developer, not a manager agent, arranges these agents into a pipeline, specifying the order of task execution and how the data flows between them. This is described as a "hierarchical process".
- LangGraph [13] is the formal abstraction of this entire model. It replaces agentic chaos with a control flow paradigm based on directed state graphs. Each agent or tool is a node in the graph, and the governance is defined by the explicit edges that create the state machine.

This Prescriptive model is a direct and intelligent engineering response to the brittleness of the Hierarchical model. The failure of Model 1 is the cognitive bottleneck of the autonomous manager. Model 2 solves this by removing the manager's autonomy. The human developer effectively becomes the manager and hard codes the entire workflow into a directed graph or SOP. This brilliantly solves the manager-brittleness problem but, in doing so, transfers the entire cognitive load of governance onto the human developer. This creates a new, more profound failure: a total inability to handle novelty. The system is now robustly prescriptive and auditable, but not inherently suited to open-ended, adaptive problems.

3.3. Model 3: Democratic (Consensus-Based) Governance

- Core Mechanism: This model governs collective decision-making, conflict resolution, and information validation. It replaces autocratic or bureaucratic dictates with social-choice mechanisms like voting, debate, and deliberation.
- Literature Analysis: This model exists almost exclusively in 2024-2025 academic research, not in production frameworks. It is emerging as the primary solution to the reasoning and conflict-resolution failures of Models 1 and 2.
- Deliberation and Debate: The ReConcile framework (ACL 2024) [9] is a prime example. It creates a round table conference for multi-round discussion where diverse LLM agents learn to convince other agents to improve their answers and reach a better consensus. This is a formal governance protocol for improving collective reasoning.
- Voting and Social Choice: A 2024 (EMNLP) paper, "An Electoral Approach to Diversify LLM-based Multi-Agent Collective Decision-Making" [10], explicitly applies social choice theory (e.g., Borda count, Instant-Runoff Voting) to enhance collaboration through ordinal preferential voting. This approach is shown to improve reasoning and provide robustness against single points of failure. Other research explores LLM Voting to simulate complex political consensus and augment digital democracy.

A purely "democratic" system for every task would be impossibly slow and burdened by communication overhead. However, this model is not a replacement for Model 2, but a critical component. The true innovation lies in hybrid systems. A Model 2 (Prescriptive) framework like LangGraph defines an efficient, bureaucratic pipeline. Its core weakness is handling exceptions. A future, robust system would run on this Model 2 pipeline until it encounters a high-uncertainty state, a logical conflict, or a task "misaligned with human intent". At that moment, the graph would trigger a "democratic" (Model 3) sub-graph a "ReConcile round table" to deliberate, vote, and resolve the conflict before returning the validated result to the prescriptive flow. This provides the efficiency of bureaucracy with the resilience of democracy.

3.4. Model 4: Economic (Market-Based) Governance

- Core Mechanism: This model leverages economic principles for dynamic coordination and task allocation. It assumes agents are self-interested and provides them with mechanisms like internal tokens, bidding, auctions, or payoff structure[s] to compete and cooperate. Governance is not a process but an incentive structure.
- Literature Analysis: This model represents a major research gap, with the first foundational papers appearing in 2025.
- Mechanism Design: A February 2025 paper from Google Research is foundational, explicitly investigating mechanism design for large language models. It proposes a simple token auction mechanism for aggregating the outputs of multiple self-interested LLMs, formally bringing economics into the agent governance stack.
- Task Allocation: An April 2025 paper directly asks how a network of LLM-based agents optimize their task allocation and tests their performance against known solutions like the Hungarian Algorithm. This signifies a critical shift from assigned tasks (Model 1) to dynamically allocated tasks (Model 4).
- Game Theory: This research connects directly to the documented challenges in defining an appropriate payoff structure and the ongoing need for refinement in the application of game theory to multi-agent systems.
- The primary challenge of this model is its primary risk: Market Failure. The governor's dilemma is most acute here. The very implicit collusion and supercompetitive pricing identified in Part 1 are the natural-emergent properties of ungoverned market-based agents.

This model is the only one that, in theory, scales governance. Models 1 and 2 require a human to design the control structure. Model 3 is too slow. Model 4, in contrast, self-organizes through price signals. However, this implies a radical re-scoping of AgentOps. An AgentOps platform for Model 4 must cease to be an engineering debugger and become a market regulator. It must include tools from economics, such as antitrust monitors to detect collusive pricing, auditors to ensure "payoff" structures align with user goals, and user-friendly mechanism-design tools as a core governance function. This moves the problem from computer science to political economy.

3.5. Model 5: Emergent (Anarchic) Governance

- Core Mechanism: This is the model of true decentralization and self-organization. There is no central control, no pre-defined graph, and no explicit auctioneer [4] Complex emergent intelligence or higher-order structure arises from agents following simple local rules and engaging in local interactions. This is swarm intelligence in its purest form.
- Literature Analysis: This is the most theoretical and futuristic model, with 2025 research focused on two questions: can this emergence be created, and can it be measured?
- Measurement: The key paper, "Emergent Coordination in Multi-Agent Language Models" (October 2025) [3], provides the measurement framework. It introduces an information-theoretic framework using Partial Information Decomposition (PID) and Time-Delayed Mutual Information (TDMI) to "quantify" and "localize" true "dynamical emergence".
- Mechanism (The Control Knob): This same paper also provides the mechanism. It demonstrates that prompt design is the control knob for this anarchy. By adding a persona to each agent or a simple theory of mind instruction (think about what other agents might do), the researchers were able to steer a mere aggregate of agents into a true "higher-order collective". This aligns with other 2025 research showing that "autonomous knowledge-driven" prompts directly affect "emergent system dynamics.
- Properties: This model is, by definition, the most robust, offering Resilience Through Redundancy and No Single Point of Failure.

This model represents the ultimate tradeoff: resilience vs. debuggability. Models 1 and 2 are popular in enterprise because they are deterministic. An AgentOps tool that offers "rewind and replay" depends on a knowable, deterministic execution trace. A Model 5 system is non-deterministic and emergent by design. Its "chaos" is its feature, not its bug. It is impossible to use a Model 1/2 "debugger" on a Model 5 "swarm." This implies that an entirely new class of governance tools is needed: not debuggers, but "information-theoretic" sensors; not orchestrators, but environmental controls and prompt-based steering mechanisms to align this emergent behavior.

4. Conclusion

This survey of the 2024-2025 research landscape reveals a field at a critical inflection point. The explosion in agent construction has outpaced the surprisingly limited research into agent governance, creating a dangerous gap between industry practice and the emerging realities of multi-agent systemic risk.

Our analysis, structured by the proposed taxonomy, demonstrates that the field is dangerously clustered.

- Industry Clustering (Models 1 & 2): Current industry and production-ready frameworks (AutoGen, CrewAI, MetaGPT, LangGraph) fall almost exclusively into the Hierarchical (Model 1) and Prescriptive (Model 2) categories. The rationale is clear: these models are deterministic, auditable, and debuggable, which are features that enterprise customers demand. However, these traps practitioners in a paradigm of brittleness and non-adaptability. These

systems cannot handle novelty.

- Academic Exploration (Models 3, 4, 5): Conversely, the most advanced 2024-2025 academic research is focused squarely on Models 3 (Democratic), 4 (Economic), and 5 (Emergent). These models promise adaptability, scalability, and resilience, the very properties required to build robust, next-generation AI.

The field is at a crossroads. Practitioners are building brittle but controllable systems. Researchers are designing adaptive but chaotic systems. The future of AI is not in any single model but in hybrid governance and the creation of a new generation of AgentOps tools capable of managing it. This survey concludes by issuing a clear roadmap for the Ethical AI community, based on the identified gaps in our taxonomy.

5. A Future Roadmap: The Urgent Call for Research

- Urgent Priority 1 (Hybridization): We call for research to "bridge the gap" by creating hybrid governance frameworks. The most promising and immediate path is the integration of Model 3 (Democratic) consensus modules as exception-handling nodes within Model 2 (Prescriptive) graph-based frameworks. This would combine the efficiency of bureaucracy (e.g., LangGraph) with the robust, conflict-resolving power of deliberation (e.g., ReConcile).
- Urgent Priority 2 (Governing Economics): We urge the "AgentOps" community to evolve beyond its 2025 focus on engineering-centric observability. A new stack of governance tools is required for Model 4. This includes market-failure monitors, collusion-detection algorithms, and user-friendly mechanism-design interfaces. The AgentOps platform in future must look less like a debugger and more like an economic regulator.
- Urgent Priority 3 (Controlling Emergence): We call for a new paradigm of "Model 5 AgentOps" that abandons deterministic debugging and instead embraces probabilistic steering. This requires maturing the information-theoretic measurement tools to quantify emergence and honing the prompt-based control knobs that can steer and align self-organizing agent swarms.

The shift from building single agents to constructing agent societies is a profound one. This new reality demands a new science of governance, one that moves beyond prompt engineering and incorporates process engineering (Model 2), social choice theory (Model 3), economic modeling (Model 4), and environmental design (Model 5). The frameworks and research identified in this survey represent the first steps "From Anarchy to Assembly." The models proposed in our taxonomy provide the map for the crucial journey ahead.

References

[1] Beyond Static Responses: Multi-Agent LLM Systems as a New Paradigm for Social Science Research - arXiv, https://arxiv.org/html/2506.01839v2
[2] Multi-Agent Collaboration Mechanisms: A Survey of LLMs - arXiv, https://arxiv.org/pdf/2501.06322?
[3] [2510.05174] Emergent Coordination in Multi-Agent Language Models - arXiv, https://arxiv.org/abs/2510.05174
[4] Enterprise Swarm Intelligence: Building Resilient Multi-Agent AI Systems, https://builder.aws.com/content/2z6EP3GKsOBO7cuo8i1WdbriRDt/enterprise-swarm-intelligence-building-resilient-multi-agent-ai-systems
[5] WHY DO MULTI-AGENT LLM SYSTEMS FAIL? - OpenReview, https://openreview.net/pdf?id=wM521FqPvI
[6] AgentOps – AI Agent Management Made Eas - AI Agents | Saastrac, https://aiagents.saastrac.com/ai-agent/swarms/
[7] Agent Tracking with AgentOps - AG2 docs, https://docs.ag2.ai/latest/docs/use-cases/notebooks/notebooks/agentchat_agentops/
[8] (PDF) Unlocking AI Creativity: A Multi-Agent Approach with CrewAI - ResearchGate, https://www.researchgate.net/publication/386306828_Unlocking_AI_Creativity_A_Multi-Agent_Approach_with_CrewAI
[9] RECONCILE: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs - ACL Anthology, https://aclanthology.org/2024.acl-long.381.pdf
[10] [2410.15168] An Electoral Approach to Diversify LLM-based Multi-Agent Collective Decision-Making - arXiv, https://arxiv.org/abs/2410.15168
[11] [2507.01413] Evaluating LLM Agent Collusion in Double Auctions - arXiv, https://arxiv.org/abs/2507.01413
[12] Leveraging LLMs for Top-Down Sector Allocation in Automated Trading - arXiv, https://arxiv.org/html/2503.09647v4
[13] Designing Cooperative Agent Architectures in 2025 - Samira Ghodratnama, https://samiranama.com/posts/Designing-Cooperative-Agent-Architectures-in-2025/.