Journal of Cyber-Physical Security and Robotics
Volume 02, Issue 3, 33-57, 2026

Research Article

Decision Rights and Accountability in Multi-Agent AI: A Governance Model for Human Override, Escalation, and Auditability

Chirag Butani
Applied Artificial Intelligence, Recruitment Technology, and Human-AI Collaboration Independent Researcher and Co-Founder & CTO, Gofer AI Canada

*Corresponding author email: mailid

Received: 01st August, 2026 Accepted: 08th August, 2026

Abstract

The growing adoption of multi-agent artificial intelligence (AI) systems is transforming organizational decision-making by enabling autonomous agents to coordinate tasks, delegate responsibilities, access external tools, and execute actions with limited human intervention. However, this increasing autonomy creates significant governance challenges concerning decision authority, human oversight, accountability, escalation, and the traceability of agent-generated outcomes. This study develops a governance model for defining decision rights and accountability in multi-agent AI environments, with particular emphasis on human override, risk-based escalation, and end-to-end auditability. Drawing on concepts from AI governance, organizational decision-rights theory, human–AI collaboration, responsible AI, and multi-agent systems, the study identifies key governance requirements for managing distributed agency and delegated authority. The proposed model establishes differentiated levels of agent decision authority, bounded delegation rules, risk-sensitive escalation mechanisms, and explicit conditions for human approval, intervention, suspension, and override. It further introduces an accountability structure that distinguishes operational, supervisory, technical, governance, and organizational responsibility across the multi-agent decision chain. To strengthen transparency and post-decision review, the model incorporates decision provenance mechanisms that record agent identities, interactions, delegations, tool usage, policy constraints, human interventions, and final outcomes. The study argues that effective governance of multi-agent AI should not eliminate autonomy but should dynamically align autonomy with decision impact, uncertainty, reversibility, and organizational risk. The proposed framework provides a foundation for organizations seeking to deploy multi-agent AI systems while preserving meaningful human authority, accountability, regulatory compliance, and auditable decision-making.

Keywords: Multi-Agent AI; Agentic AI; AI Governance; Decision Rights; Human Override; Accountability; Risk-Based Escalation; Auditability; Decision Provenance; Human–AI Oversight.

Introduction

Background and Research Context

Artificial intelligence is moving beyond systems that simply generate predictions or respond to individual prompts. A growing class of systems can now plan tasks, assign responsibilities, communicate with other agents, access external tools, and carry out actions with limited human involvement. In multi-agent artificial intelligence systems, several autonomous or semi-autonomous agents may work together toward a shared objective, with each agent performing a distinct role such as planning, information retrieval, verification, risk assessment, or execution. This structure can improve the handling of complex tasks, but it also introduces governance problems that are less pronounced in conventional AI systems.

Traditional approaches to AI governance often assume a relatively clear relationship between a human user, a technical system, and an identifiable organizational owner. Multi-agent systems make this relationship more difficult to define because a single decision may be produced through several interconnected actions. A planner agent may divide a task among specialist agents, one specialist may invoke an external service, another may evaluate the result, and an execution agent may carry out the final action. The person who initiated the process may not observe every intermediate step. As a result, decision-making authority can become distributed across several technical and human actors.

This distribution of authority has practical consequences. In a financial institution, for example, one agent might evaluate a transaction, another might assess fraud risk, and a third might authorize or initiate an operational response. In cybersecurity, agents may detect threats, prioritize incidents, isolate devices, modify access permissions, or initiate remediation procedures. Similar arrangements can arise in healthcare, recruitment, supply-chain management, software engineering, and enterprise operations. The benefits of such systems depend partly on their ability to act without requiring human approval for every routine action. However, the same autonomy may create serious problems when agents operate beyond approved authority, interpret policies differently, delegate tasks incorrectly, or make decisions that cannot easily be reversed.

Research on agentic systems has already raised concerns about the risks associated with increased autonomy, including unintended actions, weak human visibility, and difficulties in assigning responsibility when autonomous systems behave in unexpected ways (Chan et al., 2023). These concerns become more significant in multi-agent settings because system behaviour may arise from interactions among several agents rather than from the output of a single model. Hammond et al. (2025) similarly identify coordination failures, conflict, collusion, and other forms of multi-agent risk that may emerge when increasingly capable systems interact with one another.

The governance problem is therefore not limited to whether AI should be allowed to make decisions. A more difficult question concerns which decisions particular agents are permitted to make, under what conditions, and within what boundaries. This issue is closely related to the concept of decision rights. In organizational theory, decision rights specify who possesses the authority to initiate, approve, execute, or control particular decisions. Aghion and Tirole (1997) distinguish between formal authority and real authority, showing that formally assigned decision powers do not always correspond to the actor who exercises actual influence over a decision. This distinction is especially relevant to multi-agent AI because an organization may formally assign authority to a human supervisor while, in practice, much of the decision process is controlled by autonomous agents.

Effective governance must therefore define the authority of both human and artificial actors. It must also address delegation between agents, escalation when risk exceeds acceptable limits, human intervention when an agent's action becomes inappropriate, and the recording of sufficient evidence to reconstruct the decision afterward.

Problem Statement

Existing AI governance frameworks provide important principles concerning transparency, fairness, accountability, safety, and human oversight. However, many of these frameworks were developed with individual automated systems or conventional machine-learning applications in mind. They do not always provide sufficient guidance for situations in which multiple autonomous agents communicate, delegate authority, resolve disagreements, and contribute collectively to a final outcome.

The central problem is the possible separation between decision authority and accountability. An agent may be permitted to recommend a course of action but not execute it. Another agent may receive delegated authority from a coordinating agent. A third may access a tool that produces consequences outside the AI environment. When something goes wrong, it may be unclear whether responsibility belongs to the agent that proposed the action, the agent that executed it, the human who initiated the workflow, the system owner, the developer, or the organization that deployed the system.

Algorithmic accountability research has long emphasized that responsibility cannot be established merely by examining the final technical output. Accountability also depends on the institutional arrangements, procedures, and actors surrounding the system (Wieringa, 2020). Novelli et al. (2024) further explain accountability as involving relationships in which actors must be able to provide reasons for their actions and may be held responsible for resulting consequences. In multi-agent systems, this requirement becomes harder to satisfy because the path from initial instruction to final action may involve many interconnected decisions.

Human oversight creates another difficulty. It is common to assume that a human supervisor can correct an automated system when necessary. In practice, meaningful oversight requires much more than placing a human somewhere in the decision process. A supervisor must receive the right information, possess sufficient authority, understand the consequences of intervention, and have enough time to act. Parasuraman et al. (2000) demonstrated that human involvement can occur at different levels of automation, while Santoni de Sio and van den Hoven (2018) argue that meaningful human control requires appropriate links between human intentions, system behaviour, and moral responsibility. Green (2022) also cautions that requirements for human oversight can become ineffective when the human role is poorly designed or merely procedural.

These concerns point to a significant governance gap. Multi-agent AI systems require a structured mechanism that determines when agents may act autonomously, when they must seek approval, when a decision must be escalated, and when humans must be able to suspend, reverse, or terminate an action. The system must also preserve records capable of showing how authority moved through the agent network and how the final decision was produced.

Decision provenance is particularly relevant to this requirement. Singh et al. (2019) argue that accountable systems require visibility into the flow of data and decisions across complex technical environments. Similarly, Raji et al. (2020) emphasize the value of end-to-end auditing that examines the complete lifecycle of an automated system rather than isolated model outputs. For multi-agent AI, such an approach means recording not only the final action but also agent identities, instructions, delegations, tool calls, intermediate recommendations, policy checks, disagreements, human approvals, overrides, and relevant system versions.

Without these controls, organizations may gain operational autonomy while losing clarity about who is authorized to act and who must answer for the consequences.

Research Aim

This study aims to develop a governance model for the allocation of decision rights and accountability in multi-agent AI systems, with particular attention to human override, risk-based escalation, delegated authority, decision provenance, and auditability.

The proposed model is intended to address the gap between broad principles of responsible AI and the operational requirements of systems in which several autonomous agents participate in the same decision process. Rather than treating autonomy as a single system-wide characteristic, the study considers autonomy as an authority that should be assigned according to the type of decision, level of risk, uncertainty, reversibility, organizational policy, and potential impact.

The model also recognizes that human oversight should be selective rather than universal. Requiring a person to approve every routine action could reduce the practical value of multi-agent systems and create oversight fatigue. At the same time, unrestricted autonomy would be inappropriate for high-impact, uncertain, or irreversible decisions. The study therefore seeks to define conditions under which authority should remain with an agent and conditions under which control should return to a human decision-maker.

Research Objectives

The study pursues the following objectives

- To examine how decision authority can be allocated between human actors and autonomous agents in multi-agent AI systems.
- To identify governance risks arising from agent autonomy, delegation, coordination, and distributed decision-making.
- To define levels of decision rights that distinguish recommendation, approval, conditional execution, bounded autonomy, and decisions reserved for human authority.
- To establish conditions that require human review, escalation, intervention, suspension, or override.
- To develop an accountability structure capable of distinguishing operational, supervisory, technical, governance, and organizational responsibility.
- To identify the information that should be captured to support traceability, decision provenance, independent review, and auditability.
- To propose an integrated governance model that connects decision rights, delegation boundaries, risk assessment, escalation, human control, accountability, and audit records.

Research Questions

The study is guided by the following research questions

- RQ1: How should decision rights be distributed between human actors and autonomous agents in multi-agent AI systems?
- This question examines how authority should be assigned across different actors and whether particular categories of decisions should remain exclusively under human control.
- RQ2: What factors should determine whether an AI agent may act autonomously, require human approval, or escalate a decision?
- This question considers factors such as risk, uncertainty, potential harm, financial impact, legal consequences, reversibility, agent disagreement, and policy constraints.
- RQ3: How can responsibility and accountability be assigned when decisions emerge from interactions among multiple AI agents?
- The question addresses the difficulty of attributing responsibility where several agents contribute to planning, analysis, verification, delegation, and execution.
- RQ4: What human override mechanisms are required to preserve meaningful human control without removing the operational benefits of agent autonomy?
- This question examines when humans should be able to pause, reject, modify, reverse, restrict, or terminate agent actions and how these powers should be incorporated into system governance.
- RQ5: What information must be recorded to ensure traceability, explainability, and auditability of consequential multi-agent decisions?
- This question focuses on the evidence required to reconstruct a decision, including agent identities, delegation paths, data sources, tool usage, policy checks, intermediate outputs, human approvals, interventions, and final outcomes.

Together, these questions shift the governance discussion from the broad issue of whether AI should be autonomous toward a more precise concern: who has the authority to make a particular decision, how that authority

may be delegated, when human control must take precedence, and how the complete process can be examined afterward. This approach is consistent with emerging work on agentic AI governance, which emphasizes visibility, controllability, and accountability as increasingly important as AI systems gain greater capacity to act independently (Chan et al., 2024; Shavit et al., 2023).

Literature Review and Theoretical Foundations

From Conventional AI to Multi-Agent Agentic Systems

Artificial intelligence systems have evolved from narrowly defined predictive tools into systems capable of planning, reasoning, communication, tool use, and autonomous task execution. Earlier machine learning systems were largely designed to perform bounded tasks such as classification, forecasting, recommendation, or pattern recognition. Generative AI expanded this capability by allowing systems to produce text, code, images, and other forms of content in response to user instructions. More recently, agentic systems have introduced an additional layer of autonomy by enabling artificial agents to pursue goals, break tasks into smaller activities, use external tools, retain contextual information, and make sequential decisions with limited human direction.

Multi-agent AI extends this development by distributing tasks among several interacting agents. Rather than relying on a single model to complete an entire process, different agents may perform specialized functions such as planning, retrieval, verification, risk assessment, negotiation, monitoring, and execution. Wooldridge and Jennings (1995) describe intelligent agents as systems capable of autonomous action within an environment while pursuing assigned objectives. Jennings (2000) further notes that agent-based systems are particularly useful for complex environments in which tasks are distributed across interacting components. These earlier foundations remain highly relevant because contemporary multi-agent AI systems follow the same underlying principle of distributed autonomous action, although they operate with far greater linguistic, reasoning, and tool-use capabilities.

The governance implications of this development are significant. A conventional AI system may generate a recommendation that is then reviewed by a human. In contrast, a multi-agent system may allow one agent to generate a plan, another to collect information, another to assess risks, and a final agent to execute the decision. The resulting action may therefore emerge from several linked decisions rather than from a single identifiable source.

Chan et al. (2023) argue that increasingly agentic systems create new categories of risk because they can pursue goals and produce consequences through sequences of autonomous actions. These risks become more difficult to manage when several agents interact. Hammond et al. (2025) identify coordination failures, conflict, strategic behavior, and other multi-agent risks that may arise when autonomous systems operate together. Such risks indicate that governance structures developed for single-model systems may not be sufficient for multi-agent environments.

Multi-agent governance must therefore address both individual agent behavior and the relationships between agents. The important question is no longer only whether an agent behaves correctly. Governance must also determine whether an agent had the authority to take the action, whether it was permitted to delegate that authority, whether another agent should have challenged the decision, and whether a human should have been involved before the action was executed.

Decision Rights in Organizations and Information Systems

Decision rights refer to the allocation of authority over particular decisions within an organization. They determine who has the power to initiate, recommend, approve, execute, monitor, or reverse a decision. This concept provides an important theoretical foundation for governing multi-agent AI because autonomous systems increasingly perform activities that were traditionally assigned to human employees, managers, or technical administrators.

Aghion and Tirole (1997) distinguish between formal authority and real authority. Formal authority refers to the officially assigned right to make a decision, while real authority refers to the actor that effectively determines the outcome. This distinction has direct relevance to multi-agent AI. An organization may formally state that a human manager remains responsible for a decision, but the practical decision process may be dominated by agents that gather evidence, generate recommendations, rank alternatives, and execute approved actions. If the human provides only nominal oversight, formal decision rights may remain human while real decision authority has shifted toward the system.

Fama and Jensen (1983) similarly distinguish between different stages of organizational decision processes, including decision initiation, ratification, implementation, and monitoring. Applying this logic to multi-agent AI suggests that authority does not need to be concentrated in a single actor. One agent may be permitted to propose an action, another may evaluate it, and a human may retain final approval for high-impact decisions.

Information systems governance research also shows that the effectiveness of decision authority depends on how rights and responsibilities are distributed. Sambamurthy

and Zmud (1999) argue that governance arrangements are shaped by organizational contingencies and the distribution of authority between different actors. In multi-agent AI, these contingencies may include the sensitivity of the task, the financial or legal consequences of a decision, uncertainty in the available evidence, the reversibility of an action, and the level of autonomy assigned to each agent.

A useful governance model must therefore distinguish between several forms of decision right

- the right to gather information;
- the right to generate recommendations;
- the right to approve an action;
- the right to execute an action;
- the right to delegate a task;
- the right to monitor other agents;
- the right to intervene or terminate an activity.

This distinction prevents organizations from treating autonomy as a simple binary condition. An agent may have the authority to execute low-risk operational decisions while remaining prohibited from approving high-impact actions. Decision rights can therefore be assigned according to the type and consequence of the decision rather than giving an agent unrestricted authority.

Human Oversight and Meaningful Human Control

Human oversight remains one of the central principles of responsible AI governance. However, the meaning of oversight becomes less straightforward as systems become more autonomous. Human participation can take several forms, including direct approval before action, monitoring during operation, intervention when abnormalities occur, and retrospective review after decisions have been made.

Parasuraman et al. (2000) developed an influential model of human interaction with automation, showing that automation can operate at different levels across information acquisition, analysis, decision selection, and action implementation. Their work demonstrates that human involvement is not simply present or absent. The appropriate level of oversight depends on the function being automated and the risks associated with that function.

This distinction is particularly important in multi-agent systems. Requiring a person to manually approve every action would substantially reduce the efficiency gains associated with autonomous agents. At the same time, removing meaningful human involvement could allow agents to make consequential decisions without adequate review. The governance challenge is therefore to determine when human intervention is necessary.

Santoni de Sio and van den Hoven (2018) argue that meaningful human control requires more than the technical ability to stop an automated system. Human actors must maintain a meaningful connection to system behavior and remain capable of understanding and influencing decisions. In practical terms, a human supervisor should receive enough information to recognize when intervention is necessary and should have sufficient authority to act before serious consequences occur.

Green (2022) cautions that human oversight requirements can fail when they are treated merely as procedural safeguards. A human may technically approve an algorithmic decision without having enough information, expertise, or time to question the system's recommendation. This concern is relevant to multi-agent AI because the complexity of agent interactions can make oversight more difficult. A supervisor may be presented with the final recommendation without seeing how several agents reached that conclusion.

Effective oversight therefore requires mechanisms that allow humans to understand the status of an agent process, observe relevant disagreements or policy violations, and intervene when predetermined conditions are met. Human control should also include the authority to pause a workflow, reject an action, revoke permissions, reduce an agent's level of autonomy, isolate an agent, or terminate a process.

Accountability in Artificial Intelligence

Accountability refers to the obligation to explain and justify decisions and to accept responsibility for their consequences. In artificial intelligence governance, accountability becomes difficult because technical systems may participate directly in decisions while legal and organizational responsibility continues to rest with human institutions.

Diakopoulos (2016) argues that algorithmic accountability requires mechanisms through which the behavior of algorithmic systems can be examined and questioned. This includes understanding how decisions are made, what information influences them, and which individuals or institutions are responsible for their operation. Wieringa (2020) similarly emphasizes that algorithmic accountability should be examined within its broader social and organizational context rather than being reduced to technical explainability.

Novelli et al. (2024) describe accountability as a structured relationship involving responsibility, answerability, and the possibility of consequences when duties are not fulfilled. This distinction is particularly useful for multi-agent AI because different actors may hold different forms of responsibility. A developer may be responsible for technical design, a system owner may be accountable for deployment, a human supervisor may be responsible for reviewing escalated cases, and an execution agent may be responsible for carrying out an authorized action. The difficulty arises when the final outcome cannot be

attributed to a single agent or person. For example, a planner agent may produce an inappropriate strategy based on information supplied by another agent, while a verifier agent fails to identify the problem and an execution agent carries out the resulting action. Assigning all responsibility to the executing agent would overlook the distributed nature of the decision.

A governance framework for multi-agent systems should therefore distinguish between at least five forms of responsibility

- Operational responsibility, relating to the actor performing the immediate task.
- Supervisory responsibility, relating to the person or system responsible for monitoring the process.
- Technical responsibility, relating to the design, configuration, and maintenance of the system.
- Governance responsibility, relating to policies, authority boundaries, and control structures.
- Organizational accountability, relating to the entity ultimately answerable for the deployment and consequences of the system.

This approach prevents accountability from disappearing simply because several agents contributed to a decision.

Delegation and Distributed Agency

Delegation is a defining feature of many multi-agent systems. A coordinating agent may assign subtasks to specialist agents, which may in turn interact with tools, services, or other agents. Delegation can improve efficiency because agents with specialized capabilities can handle different parts of a complex workflow. However, it also creates a risk that authority may be transferred beyond its intended limits.

A typical decision chain might take the following form

Human → Planner Agent → Specialist Agent → External Tool → Verification Agent → Execution Agent

Each transfer creates a governance question. The planner agent may be authorized to coordinate work but not to authorize a financial transaction. A specialist agent may be permitted to access a database but not to share sensitive information with another system. An execution agent may have technical access to an external tool even when policy rules prohibit a particular action.

Chan et al. (2024) emphasize the importance of visibility into AI agents because governance becomes difficult when organizations cannot clearly observe what agents are doing, what resources they are using, and how their actions relate to one another. Visibility is especially important in delegated workflows where responsibility can become separated from execution.

Delegation rules should therefore define what authority can be transferred and under what conditions. A central principle is that an agent should not be permitted to delegate more authority than it possesses. If an agent has permission to recommend a transaction but not approve it, it should not be able to assign approval authority to another agent.

Delegation may also require limits relating to

- duration of authority;
- financial or operational value;
- data access;
- tool permissions;
- task scope;
- number of delegation layers;
- permitted recipient agents;
- ability to subdelegate.

These controls are necessary because recursive delegation can make it difficult to determine where authority originated and who remains responsible for the final outcome.

Explainability, Traceability, and Auditability

Explainability, traceability, provenance, and auditability are related concepts, but they serve different governance purposes.

Explainability concerns the reasons for a decision. It attempts to answer the question: Why did the system reach this conclusion? Traceability addresses the sequence of events that produced the outcome. It asks: What happened, and in what order? Provenance concerns the origin and movement of data, instructions, decisions, and actions across the system. Auditability concerns whether an independent reviewer can reconstruct and evaluate the process.

These distinctions are essential in multi-agent AI because understanding a final model output may not reveal how the overall decision occurred. A reviewer may need to determine which agent generated a recommendation, what information it used, whether another agent challenged the recommendation, which policy was applied, whether a human approved the action, and which tool finally executed it.

Singh et al. (2019) introduce the concept of decision provenance as a means of tracing how data and decisions move across interconnected systems. Their approach is especially suitable for multi-agent AI because accountability often depends on understanding relationships between multiple components rather than examining an isolated model output.

Raji et al. (2020) similarly argue for end-to-end algorithmic auditing. Their framework emphasizes that auditing should cover the entire lifecycle and organizational process surrounding an automated system. Applied to multi-agent systems, this means that audit records should include more than prompts and final responses.

Relevant records may include

- agent identities;
- model versions;

Table 1: Existing AI Governance Concepts and Their Limitations for Multi-Agent AI

Governance Area	*Primary Question*	*Traditional Approach*	*Multi-Agent Governance Challenge*	*Requirement for Proposed Model*
Decision rights	Who may decide?	Authority assigned to human or system roles	Several agents may influence or execute one decision	Explicit agent-level decision authority
Delegation	Who may transfer authority?	Human-to-human or human-to-system delegation	Agents may delegate tasks and authority to other agents	Bounded and traceable delegation rules
Human oversight	When should humans intervene?	Human-in-the-loop or monitoring arrangements	Decisions may occur rapidly across several agents	Risk-based and timely intervention mechanisms
Accountability	Who is responsible for the outcome?	Responsibility linked to developers, operators, or institutions	Responsibility may be distributed across many agents and humans	Clear operational, supervisory, technical, and organizational accountability
Explainability	Why was the decision made?	Explanation of a model's output	The final outcome may result from several models and interactions	End-to-end explanation of the decision process
Escalation	When should approval be required?	Static thresholds or manual approval	Risk may change during an agent workflow	Context-sensitive escalation rules
Auditability	Can the decision be independently reviewed?	System logs and model documentation	Complex agent interactions create fragmented evidence	Complete and tamper-resistant decision provenance

- system instructions;
- data sources;
- delegated tasks;
- tool and API calls;
- intermediate recommendations;
- confidence or uncertainty measures;
- disagreements between agents;
- policy checks;
- escalation events;
- human approvals;
- override actions;
- final execution details;
- resulting outcomes.

Mökander et al. (2021) further explain that auditing can strengthen accountability by providing a structured means of evaluating whether automated systems operate according to ethical and governance requirements. In multi-agent environments, auditability therefore depends on preserving a complete and reliable record of how authority and information moved through the system.

Research Gap

The literature provides substantial guidance on AI ethics, accountability, human oversight, organizational authority, autonomous agents, and algorithmic auditing. However, these areas are often examined separately. Research on organizational decision rights primarily concerns human actors and institutional structures. Human oversight research generally focuses on interactions between people and individual automated systems. Accountability studies frequently examine algorithms or organizational processes without addressing the complexity created by multiple cooperating autonomous agents.

Recent work on agentic systems has begun to identify risks associated with greater autonomy and distributed interaction (Chan et al., 2023; Chan et al., 2024; Hammond et al., 2025). However, there remains a need for a governance model that translates these concerns into practical rules for allocating authority within multi-agent systems.

In particular, limited attention has been given to how organizations should determine

- which agents may make particular decisions;
- how authority may be delegated between agents;
- what circumstances should trigger human escalation;
- when humans must be able to override or terminate agent actions;
- how accountability should be distributed across several contributors;
- what evidence must be preserved to reconstruct the complete decision process.

This gap is important because autonomy, accountability, and auditability cannot be treated as independent governance issues. Greater autonomy changes the distribution of decision rights. Delegation changes the location of responsibility. Escalation determines when control returns to humans. Auditability determines

whether those relationships can be examined after the event.

The present study addresses this gap by integrating these elements into a single governance structure

Decision Rights → Delegation → Risk Assessment → Escalation → Human Override → Accountability → Auditability

The proposed approach treats decision rights as the starting point of governance rather than relying mainly on retrospective accountability. It defines what agents are permitted to do before a decision is executed, identifies conditions under which authority must be restricted or transferred to a human, and establishes an evidence trail that allows organizations to determine how consequential outcomes were produced.

The literature suggests that multi-agent AI governance requires more than general principles of transparency and human oversight. It requires explicit rules defining who can decide, who can delegate, when autonomy must stop, who can intervene, who remains accountable, and what evidence must exist to verify the process afterward. This theoretical foundation informs the decision-rights and accountability model developed in the following sections.

Research Methodology

Research Design

This study adopts a conceptual framework development approach supported by structured literature analysis and scenario-based evaluation. The purpose of the methodology is not to test a single algorithm or compare model performance, but to develop a governance structure that can guide the allocation of decision authority, human intervention, accountability, and auditability in multi-agent AI systems.

A conceptual approach is appropriate because the governance of multi-agent AI remains an emerging area in which technical capabilities are developing faster than organizational control mechanisms. Existing studies provide useful insights into autonomous agents, algorithmic accountability, human oversight, decision provenance, and AI governance, but these areas are often treated separately. The present study brings these strands together to establish a coherent governance model for multi-agent decision-making.

The research design is informed by work on algorithmic accountability and auditing, particularly the argument that governance should examine the full lifecycle of an automated system rather than focusing only on its final output. Raji et al. (2020) emphasize the importance of end-to-end accountability processes that connect technical development, organizational practices, documentation, and review. This perspective is especially relevant to multi-agent AI because a final action may result from several interconnected agents, tools, and human interventions.

The study proceeds through four linked stages:

structured review of relevant literature;

identification of governance requirements;

development of the proposed decision-rights and accountability model;

scenario-based evaluation of the model.

This sequence allows the framework to be grounded in established literature while remaining applicable to contemporary multi-agent AI environments.

Literature Identification and Selection

The first stage involves a structured review of literature relating to multi-agent systems, AI governance, organizational decision rights, human oversight, algorithmic accountability, autonomous systems, and auditability. The literature base includes peer-reviewed journal articles, conference papers, foundational theoretical studies, and selected governance reports that directly address the design, control, supervision, and accountability of autonomous or semi-autonomous systems.

Relevant studies may be identified from major academic databases such as Scopus, Web of Science, IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, and Google Scholar. Search terms should combine concepts from both technical and governance domains, including

- multi-agent systems;
- agentic AI;
- autonomous agents;
- AI governance;
- decision rights;
- delegated authority;
- human oversight;
- meaningful human control;
- algorithmic accountability;
- AI auditing;
- decision provenance;
- traceability;
- escalation;
- human override.

The review gives particular attention to literature that explains how authority is distributed, how automated actions are supervised, and how decisions can be reconstructed after execution. Foundational studies on organizational authority are also included because decision rights cannot be understood solely as a technical problem. Aghion and Tirole (1997), for example, distinguish formal authority from real authority, a distinction that is highly relevant when organizations retain nominal human control while agents exercise significant practical influence over decisions.

The literature is not treated merely as background

material. Instead, it is examined for specific governance mechanisms, limitations, and design requirements that can be translated into a multi-agent context.

Analytical Dimensions

The selected literature is analyzed across a set of governance dimensions that reflect the main challenges associated with multi-agent decision-making. These dimensions are used to compare existing approaches and identify areas where current governance models remain insufficient.

The primary dimensions are

- Agent autonomy, referring to the degree to which an agent can act without direct human approval.
- Decision authority, referring to the type of decisions an agent or human is permitted to make.
- Delegation, referring to the transfer of tasks or authority between agents.
- Human oversight, referring to the extent and timing of human supervision.
- Risk classification, referring to the assessment of potential harm, uncertainty, impact, and reversibility.
- Escalation, referring to the conditions under which a decision must be transferred to a higher authority.
- Override capability, referring to the ability of authorized humans to pause, modify, reverse, or terminate an agent action.
- Accountability, referring to the assignment of responsibility for decisions and outcomes.
- Explainability, referring to the ability to provide understandable reasons for a decision.
- Traceability, referring to the ability to reconstruct the sequence of actions and interactions.
- Auditability, referring to the ability of an independent reviewer to evaluate whether the decision process complied with established rules.

These dimensions are particularly important because multi-agent AI introduces governance relationships that are not present in many conventional automated systems. An agent may be technically capable of taking an action while lacking the organizational authority to do so. Similarly, an agent may have permission to perform a task but not permission to delegate that task to another agent.

Framework Development

The proposed governance model is developed through a synthesis of the findings from the literature review. The framework is structured around the principle that autonomy should be treated as a controlled form of decision authority rather than an unrestricted technical capability.

The development process consists of four stages

Literature synthesis → Governance requirement identification → Model construction → Scenario-based evaluation

During literature synthesis, the study identifies established principles relating to accountability, human control, decision authority, provenance, and auditing. These principles are then translated into operational governance requirements.

For example, research on meaningful human control suggests that humans must retain genuine influence over consequential autonomous actions rather than merely being included as symbolic reviewers (Santoni de Sio & van den Hoven, 2018). Research on decision provenance emphasizes the importance of tracing how information and decisions move across technical systems (Singh et al., 2019). Accountability literature further shows that responsibility should be linked to organizational processes, not simply to the final algorithmic output (Wieringa, 2020; Novelli et al., 2024).

These insights are combined into a model that includes

- explicit decision-right levels;
- bounded delegation;
- risk-based escalation;
- human override;
- accountability assignment;
- decision provenance;
- audit requirements.

The model is designed to govern both routine autonomous actions and high-impact decisions that require human intervention.

Scenario-Based Evaluation

The final methodological stage evaluates the proposed model through representative multi-agent scenarios. Scenario-based evaluation is used because the governance challenges examined in this study depend heavily on context. The same degree of autonomy may be acceptable for one task but inappropriate for another.

The scenarios may include

- autonomous financial transaction processing;
- cybersecurity incident response;
- automated recruitment screening;
- supply-chain procurement;
- autonomous software deployment.

Each scenario is evaluated using a common set of governance questions

- Which agent or human holds the relevant decision right?
- Is the proposed action within the agent's authority?
- Has any authority been delegated?
- Is the delegation permitted?
- What level of risk is associated with the action?
- Does the decision require escalation?
- Is human approval necessary?

- Can the action be reversed?
- Who is accountable for the final outcome?
- What records are necessary for audit and review?

The purpose of this evaluation is not to claim empirical validation across all industries. Instead, it tests whether the proposed framework provides consistent and interpretable governance decisions across different operational settings.

Decision Rights in Multi-Agent AI

The Decision-Rights Problem

Decision rights determine who is authorized to make a particular decision and what type of action that authority includes. In conventional organizations, decision rights are usually assigned to individuals, teams, managers, or committees. In multi-agent AI environments, part of this authority may be transferred to artificial agents that can analyze information, recommend actions, delegate tasks, and execute decisions.

The central governance problem is that technical capability does not automatically create legitimate authority. An agent may have access to a database, payment system, security tool, or software environment, but access alone does not mean that the agent should be authorized to use that capability in every circumstance. Aghion and Tirole's (1997) distinction between formal and real authority is useful in this context. An organization may formally state that a human remains responsible for a decision, while the AI system effectively determines the available options, ranks them, and carries out the action. If human involvement becomes largely procedural, the real decision authority may rest with the system even when formal responsibility remains with a person.

Multi-agent AI further complicates the issue because authority may be distributed across several agents. One agent may be authorized to analyze a problem, another to recommend an action, another to verify compliance, and another to execute the final decision. Governance therefore requires a clear definition of the authority assigned to each participant.

A useful concept is the authority envelope*, which defines the limits within which an agent may operate. The authority envelope should specify*

- the decisions the agent may make;
- the actions the agent may execute;
- the systems and tools it may access;
- the data it may use;
- the financial or operational limits of its actions;
- the agents to which it may delegate tasks;
- the conditions under which it must seek approval;
- the actions it is explicitly prohibited from taking.

An agent acting outside this envelope should not merely generate a warning. The action should be blocked or escalated according to the seriousness of the violation.

Proposed Decision-Authority Levels

A practical governance model requires more than a distinction between autonomous and non-autonomous systems. Decision authority should be divided into levels so that agents can receive different degrees of autonomy depending on risk, context, and organizational policy. The following five levels are proposed.

Level 0: Observe

- At this level, the agent may collect, retrieve, summarize, or organize information but cannot make recommendations or execute decisions.
- This level is suitable for highly sensitive environments where the agent functions primarily as an information support tool.

Examples include

- retrieving compliance records;
- monitoring system status;
- summarizing operational data;
- collecting evidence for human review.

Level 1: Recommend

The agent may analyze information and propose a course of action, but the final decision remains with a human. The agent may rank alternatives, identify risks, or suggest responses, but it cannot independently execute the recommendation.

This level is appropriate for decisions involving

- significant financial consequences;
- legal implications;
- personnel decisions;
- regulatory interpretation;
- sensitive organizational judgment.
- The relationship resembles decision support rather than autonomous decision-making.

Level 2: Conditional Execution

- At this level, the agent may prepare and execute an action only after explicit authorization is provided.
- For example, an agent may draft a transaction, security response, procurement request, or software change, but execution remains blocked until an authorized human or governance function approves it.
- This arrangement is useful where automation is valuable but the consequences justify a final approval checkpoint.

Level 3: Bounded Autonomy

The agent may act independently within clearly defined limits.

These limits may include

- transaction value;
- data sensitivity;
- geographical scope;
- tool permissions;
- operational risk;
- time period;
- permitted counterparties;
- predefined policy conditions.

This level is likely to be the most common in practical multi-agent systems because it allows organizations to gain operational efficiency while preserving governance boundaries.

For example, a cybersecurity agent may automatically block suspicious network traffic when confidence is high and the action is reversible, but it may require human approval before disabling a critical production system.

Level 4: High Autonomy with Human Oversight

At this level, the agent may execute complex decisions without prior approval, provided that its activities are continuously monitored and remain subject to intervention.

The human role changes from direct approval to supervisory control.

Human supervisors should receive

- alerts for abnormal behavior;
- high-risk decisions;
- policy violations;
- agent disagreement;
- unusual tool use;
- repeated failures;
- unexpected escalation patterns.

This form of autonomy should only be used where the organization has reliable monitoring, strong audit controls, and effective override mechanisms.

Level 5: Human-Reserved Decision

Certain decisions should not be delegated to autonomous agents regardless of technical capability.

These may include decisions involving

- severe safety consequences;
- irreversible legal effects;
- termination of critical services;
- highly sensitive personnel actions;
- major financial exposure;
- significant regulatory consequences.

In these cases, agents may provide analysis and recommendations, but final authority remains with an authorized human.

This level reflects the principle that some forms of organizational authority should remain explicitly human.

Delegation Rights

Multi-agent systems depend heavily on delegation. A planner agent may assign tasks to retrieval agents, verification agents, risk agents, or execution agents. Delegation can improve efficiency, but it also creates one of the most important governance risks in multi-agent systems.

The central rule proposed in this study is

- An agent must not delegate authority that it does not possess.
- If an agent is authorized only to recommend an action, it should not be able to delegate execution authority to another agent. Likewise, if an agent has permission to access a limited category of data, it should not be permitted to give another agent broader access.

Delegation should therefore be governed by explicit rules covering:

- scope;
- duration;
- purpose;
- financial limit;
- data-access level;
- tool permissions;
- recipient identity;
- ability to subdelegate;
- escalation requirements.

A delegation record should also identify the source of authority. This creates a traceable chain showing where permission originated and how it moved through the system.

For example

Human Supervisor → Planner Agent → Risk Agent → Execution Agent

Each transfer should preserve the original authority limits. The execution agent should never receive more authority than was granted to the planner.

This principle becomes especially important when agents can dynamically create tasks or call external services. Without delegation controls, authority can gradually expand across the system without deliberate organizational approval.

Chan et al. (2024) emphasize the importance of visibility into agent behavior, particularly as autonomous systems interact with tools and other agents. In governance terms, visibility should include not only what an agent did but also why it was allowed to do it and from whom that authority was derived.

Dynamic Decision Rights

Decision rights should not always remain fixed. The appropriate level of autonomy may change as the context of a decision changes.

An agent that normally operates with bounded autonomy

may need to surrender authority when uncertainty or potential impact increases. Conversely, routine low-risk actions may not require repeated human approval if they occur within well-defined limits.

Dynamic decision rights can therefore be adjusted using factors such as

- decision risk;
- confidence level;
- agent disagreement;
- financial exposure;
- safety consequences;
- reversibility;
- policy compliance;
- data sensitivity;
- novelty of the situation;
- historical agent performance.

A simple governance relationship can be expressed as

As risk or uncertainty increases, the level of autonomous authority should decrease.

For example, an agent may normally have Level 3 bounded autonomy. If its confidence falls below an approved threshold or another agent raises a serious objection, its authority may automatically shift to Level 2, requiring human approval before execution.

Similarly, an irreversible action should normally attract stricter authority requirements than a reversible action with limited operational impact.

This approach is consistent with the broader argument that human oversight should be proportionate to the consequences of automation rather than applied uniformly. Parasuraman et al. (2000) show that different levels of automation require different forms of human involvement, while Green (2022) cautions that oversight becomes ineffective when human participation is poorly matched to the actual decision context.

Dynamic decision rights therefore provide a practical middle ground between two problematic extremes. One extreme is unrestricted autonomous action. The other is constant human approval of every minor decision. A risk-sensitive model allows autonomy where it is useful while returning authority to humans when the consequences become more serious.

Taken together, the proposed decision-right structure establishes a clear rule for multi-agent governance: autonomous capability should never be assumed to equal organizational authority. Every agent should operate within a defined authority envelope, every delegation should remain bounded, and decision rights should contract when risk, uncertainty, or potential harm increases.

Proposed Decision Rights and Accountability Governance Model

The proposed Decision Rights and Accountability Governance Model is designed to provide a practical structure for controlling authority, supervision, escalation, accountability, and auditability in multi-agent AI systems. The model responds to a central problem in agentic environments: several agents may contribute to a single outcome, but their authority, responsibilities, and limits are not always clearly defined. Without explicit governance controls, organizations may know what an agent is technically capable of doing while remaining uncertain about what it is actually permitted to do.

The model is built around the principle that decision authority must be explicitly assigned, bounded, monitored, and reviewable. Each participating agent should operate within an approved authority envelope that specifies the decisions it may make, the tools it may use, the data it may access, the actions it may execute, and the conditions that require human intervention. This separates technical capability from organizational authority and reduces the risk that agents perform actions simply because system access allows them to do so.

At the top of the model is the organizational governance and policy layer. This layer establishes the rules under which multi-agent systems operate. It defines risk tolerance, prohibited actions, regulatory obligations, decision classes, approval requirements, accountability ownership, and audit requirements. These policies provide the basis for determining what agents are permitted to do and when control must return to a human authority. Novelli et al. (2024) emphasize that accountability depends not only on system behavior but also on clearly defined relationships of responsibility and answerability. For this reason, organizational governance should remain the ultimate source of decision authority.

Below this layer is the human authority layer, which identifies the individuals or organizational functions that retain supervisory and final decision rights. Human authority may include system owners, compliance officers, risk managers, business owners, security personnel, or other designated decision-makers. Their role is not necessarily to approve every action. Instead, they provide governance oversight, resolve high-risk situations, authorize exceptional actions, and intervene when agent behavior exceeds acceptable limits.

The next component is the decision rights and authority engine. This component determines what level of authority applies to a particular decision. It evaluates the agent's assigned role, policy constraints, decision category, risk level, available evidence, sensitivity of the data involved, and potential consequences. It then determines whether the agent may proceed autonomously, must operate within specified limits, or must obtain human approval.

The authority engine should verify several conditions before allowing an action

- whether the agent has the required decision right;
- whether the action is within its approved scope;
- whether the necessary data and tools are authorized;
- whether delegation has occurred;
- whether the delegated authority remains valid;
- whether the decision exceeds financial, operational, or legal thresholds;
- whether a human approval requirement applies.
- This prevents the system from treating permission as a permanent or unrestricted capability.

The multi-agent coordination layer sits beneath the authority engine. This layer contains the agents that perform the actual work. A typical arrangement may include a planner agent, specialist agents, a verifier or critic agent, and an execution agent. The planner agent may break a task into smaller activities. Specialist agents may perform analysis, retrieval, forecasting, compliance checking, or risk assessment. A verifier agent may review the proposed decision for consistency, policy compliance, or obvious errors. The execution agent may then carry out the approved action.

The governance value of separating these roles is that authority and accountability can be assigned more precisely. A planner may have the right to coordinate tasks but not to execute financial transactions. A retrieval agent may have permission to access records but not modify them. A verifier may challenge or reject a proposed action but may not independently implement an alternative. The execution agent may perform an action only when the required authorization conditions have been met.

Chan et al. (2024) argue that visibility into agent activities is essential for effective governance because autonomous systems may act through complex interactions with tools and other agents. The proposed model therefore requires that each stage of multi-agent coordination be observable and attributable to identifiable agents.

Following multi-agent coordination, the system enters a risk and confidence assessment layer. This layer evaluates whether the proposed action remains suitable for autonomous execution. The assessment may consider

- decision impact;
- uncertainty;
- confidence;
- reversibility;
- financial exposure;
- regulatory implications;
- data sensitivity;
- agent disagreement;
- operational consequences;
- policy compliance.
- The result determines one of three principal governance paths.

The first path is autonomous execution. This is appropriate when the action falls within the agent's approved authority, risk remains low, policy requirements are satisfied, and the consequences are limited or reversible.

The second path is human escalation or approval. This path is activated when the decision exceeds predetermined thresholds, involves uncertainty, creates significant financial or regulatory exposure, or falls outside the normal operating conditions of the agent.

The third path is human override or termination. This occurs when the system detects policy violations, dangerous behavior, unauthorized activity, severe disagreement between agents, or other conditions that justify immediate intervention.

After execution, the complete process is recorded through an immutable provenance and audit layer. The purpose of this layer is to preserve the evidence required to reconstruct how the decision occurred. Singh et al. (2019) demonstrate the importance of decision provenance for understanding how information and actions move across interconnected systems. In a multi-agent environment, such records should capture not only the final output but also the sequence of contributing actions.

The audit record should include

- initiating request or event;
- agents involved;
- agent roles and versions;
- data sources consulted;
- tool and API calls;
- delegated permissions;
- intermediate recommendations;
- disagreements between agents;
- risk assessments;
- policy checks;
- escalation events;
- human approvals;
- override actions;
- final decision;
- execution details;
- resulting outcome.

Raji et al. (2020) similarly argue that accountability should be supported by end-to-end auditing rather than isolated inspection of model outputs. This principle is central to the proposed model because multi-agent decisions may involve several technical and human contributors.

The final component is the accountability review and governance feedback layer. This layer is used after important decisions, incidents, or policy exceptions. It examines whether authority was properly assigned, whether escalation occurred at the correct point, whether human intervention was effective, and whether the final

outcome complied with organizational requirements. Findings from this review can then be used to modify policy thresholds, authority levels, escalation rules, and monitoring arrangements.

The model therefore creates a continuous governance cycle

Policy → Authority Assignment → Agent Coordination → Risk Assessment → Execution or Escalation → Audit → Accountability Review → Policy Improvement

This structure allows organizations to benefit from multi-agent autonomy while preserving control over consequential decisions.

Human Override and Escalation Mechanisms

Human override and escalation are essential safeguards in multi-agent AI because autonomous authority should not continue when the circumstances surrounding a decision materially change. A system that can act independently must also contain mechanisms that allow authorized humans to regain control when risk, uncertainty, or policy violations exceed acceptable limits. Human override refers to the ability of an authorized person to interrupt, modify, reject, suspend, reverse, or terminate an agent's action. It should not be treated as a symbolic governance feature. The override mechanism must be technically effective, available at the appropriate stage of the workflow, and supported by sufficient information for the human decision-maker to understand why intervention may be necessary.

Santoni de Sio and van den Hoven (2018) argue that meaningful human control requires a genuine relationship between human intentions and system behavior. In practical terms, this means that humans must have more than nominal supervisory authority. They must be able to influence consequential actions before unacceptable outcomes become unavoidable.

The proposed governance model identifies several forms of human override.

First, a human supervisor should be able to pause an active workflow. This is important where additional information is required or where the system displays unusual behavior.

Second, authorized users should be able to reject an agent recommendation before execution. This is particularly important for high-impact decisions in finance, employment, security, healthcare, or other regulated settings.

Third, humans should be able to cancel pending actions that have been prepared but not yet executed.

Fourth, the system should support reversal of reversible actions where technically possible. For example, temporary account restrictions, access permissions, or configuration changes may be restored following human review.

Fifth, supervisors should be able to revoke agent permissions. If an agent behaves unexpectedly or attempts to exceed its assigned authority, access to tools, data, or delegated functions should be removed.

Sixth, organizations should have the ability to reduce an agent's autonomy level. An agent operating under bounded autonomy may be temporarily moved to a recommendation-only or approval-required state.

Seventh, a problematic agent may need to be isolated from the multi-agent system to prevent further delegation, communication, or tool access.

Finally, critical incidents may require termination of the complete agent workflow.

The effectiveness of these mechanisms depends on timely escalation. Escalation is the process through which a decision moves from autonomous handling to higher authority when predetermined conditions are met. The purpose is not to eliminate autonomy but to ensure that the level of human involvement increases as the seriousness of the decision increases.

A basic escalation principle can be expressed as

However, risk should not be interpreted as a single variable. A decision may require escalation because of several interacting factors even when no individual indicator appears extreme.

The proposed model therefore considers escalation triggers such as

- low agent confidence;
- high uncertainty;
- high financial exposure;
- safety consequences;
- legal or regulatory implications;
- use of sensitive data;
- conflicting recommendations between agents;
- policy violations;
- unusual tool access;
- actions outside normal operating patterns;
- irreversible consequences;
- repeated system failures;
- attempts to exceed delegated authority.

Agent disagreement is particularly important. If a planner agent recommends an action while a verifier or risk agent identifies a significant concern, the disagreement should not simply be resolved through majority voting or agent hierarchy. Depending on the severity of the conflict, the system should either request additional analysis or escalate the decision to a human authority.

The proposed framework classifies escalation according to four risk categories

- Low-risk decisions may be executed autonomously when they fall within approved authority and are easily reversible.
- Moderate-risk decisions may be executed with active

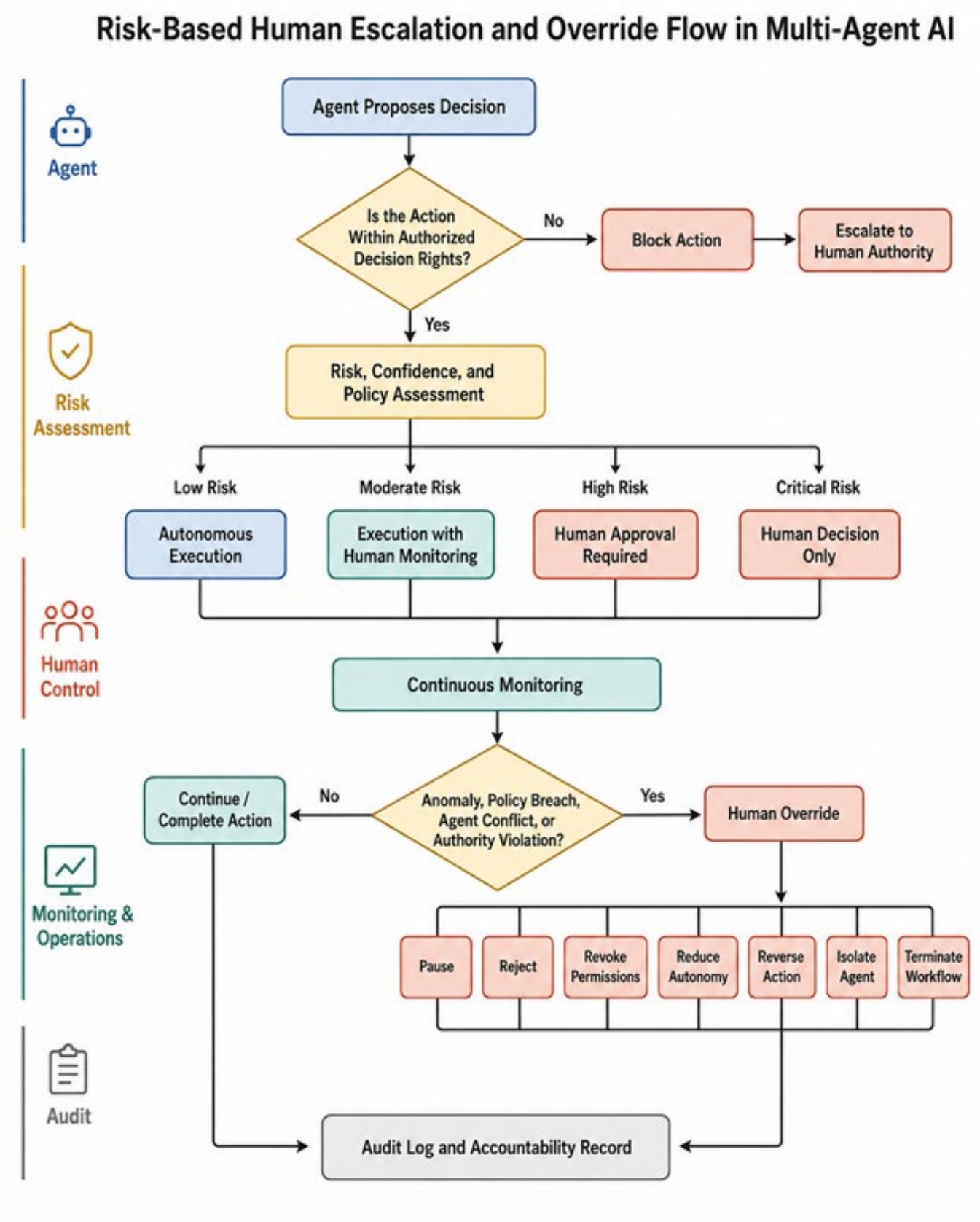

Fig 1: Risk-Based Human Escalation and Override Flow for Multi-Agent AI Governance.

monitoring. A human supervisor should be informed and able to intervene if conditions change.
- High-risk decisions should require explicit human approval before execution.
- Critical-risk decisions should remain human-reserved. Agents may provide evidence, recommendations, and analysis, but they should not possess final execution authority.

This approach reflects the principle that human involvement should be proportional to decision impact. Parasuraman et al. (2000) show that automation can be distributed across different stages of decision-making rather than applied uniformly. In a multi-agent environment, this means that human control can be concentrated where its value is greatest.

Escalation mechanisms must also prevent human oversight overload. If every minor deviation produces an alert, supervisors may become desensitized and fail to respond effectively to serious events. Escalation thresholds should therefore distinguish routine exceptions from consequential anomalies.

Green (2022) warns that human oversight can become ineffective when humans are expected to approve automated decisions without sufficient information or practical ability to challenge them. For this reason, an escalation request should include concise but adequate information such as

- proposed action;
- agents involved;
- reason for escalation;
- estimated risk;
- relevant policy;
- key evidence;
- known uncertainties;
- agent disagreements;
- consequences of approval or rejection.

The human supervisor should not be required to reconstruct the full process manually before making every decision. Instead, the system should present the most relevant governance information while preserving the complete record for audit.

Human intervention must also be logged. Every approval, rejection, modification, permission change, override, and termination should become part of the decision provenance record. This protects both organizational accountability and the integrity of subsequent audits.

The relationship between autonomy and escalation can therefore be summarized as follows

Low risk + high confidence + reversible action = greater autonomous authority

High risk + uncertainty + irreversible consequences = greater human control

This structure avoids two extremes: unrestricted agent autonomy and constant human approval of routine operations. It creates a graduated governance mechanism in which authority changes as the context and risk of a decision change.

Accountability Allocation in Multi-Agent Decisions

Moving Beyond Single-Agent Accountability

Accountability becomes more difficult when decisions are produced through interactions among several autonomous or semi-autonomous agents. In a conventional automated system, the decision pathway may involve a single model, a human operator, and an identifiable organizational owner. In multi-agent AI, however, planning, information retrieval, verification, risk assessment, tool use, and execution may be distributed across different agents. The final action can therefore result from a chain of contributions rather than the decision of one identifiable actor.

This distributed structure creates an important governance problem. If an undesirable outcome occurs, responsibility cannot automatically be assigned to the agent that executed the final action. The execution agent may have acted on instructions generated by a planner agent, relied on information retrieved by another agent, or followed a recommendation that should have been rejected by a verifier. The source of failure may therefore appear earlier in the decision process.

Algorithmic accountability requires more than identifying the technical component that produced the final output. Diakopoulos (2016) argues that accountable automated systems must provide mechanisms for examining how decisions were produced and who is responsible for them. Wieringa (2020) similarly emphasizes that accountability must consider the wider organizational, institutional, and technical environment surrounding algorithmic decision-making. These principles are especially relevant to multi-agent systems because decision-making authority may be spread across several agents and human actors.

A useful distinction can be made between five forms of responsibility. Operational responsibility concerns the agent or person carrying out a specific task. Supervisory responsibility relates to the individual or organizational function expected to monitor the system and respond when intervention is required. Technical responsibility covers the design, configuration, maintenance, access controls, and reliability of the system. Governance responsibility concerns the rules defining agent authority, delegation limits, escalation requirements, and permitted actions. Finally, organizational accountability rests with the institution that authorizes the deployment and benefits from the use of the system.

This distinction is important because autonomous agents may perform tasks, but they do not remove the responsibility of the organization that establishes the conditions under which those tasks are performed.

An organization should therefore remain answerable for the decision to grant autonomy, establish authority boundaries, select monitoring arrangements, and determine when human involvement is required.

Novelli et al. (2024) describe accountability as involving responsibility, answerability, and the possibility of consequences when obligations are not fulfilled. In multi-agent AI, this requires a governance arrangement capable of connecting each significant action to an identifiable role and authority. Accountability should therefore follow the full decision process rather than being assigned only after the final outcome has occurred.

A further challenge arises when formal responsibility does not match practical control. Aghion and Tirole (1997) distinguish between formal authority and real authority, showing that the actor officially responsible for a decision may not always be the actor exercising the greatest influence over it. In multi-agent AI, a human supervisor may formally retain decision authority while agents determine most of the analysis, ranking, verification, and recommended course of action. If the human has limited visibility or only approves decisions routinely, accountability becomes difficult to justify.

For accountability to remain meaningful, responsibility must therefore be aligned with actual decision rights, access to information, and the ability to intervene.

Accountability Chain

An accountability chain provides a structured way of identifying how responsibility moves through a multi-agent decision process. Rather than focusing only on the final action, it records who initiated the process, which agents participated, what authority they possessed, how tasks were delegated, and who approved or executed the final action.

A typical accountability chain may involve the following actors

Human or Organizational Initiator → Planner Agent → Specialist Agents → Verifier Agent → Decision Gateway → Human or Authorized Agent → Execution Agent → Outcome

The initiator establishes the original objective or task. The planner agent determines how the objective should be approached and may assign subtasks to specialist agents. These agents may retrieve information, perform analysis, assess risk, or interact with external systems. A verifier agent may then examine the recommendations for consistency, policy compliance, or reliability. Depending on the decision rights assigned to the system, the resulting action may either proceed autonomously or require human authorization before execution.

Each stage should be associated with a clearly defined responsibility. The planner agent should be accountable for operating within its planning authority. Specialist agents should be restricted to approved functions and data. Verification agents should be expected to identify relevant conflicts or policy violations. Human supervisors should be responsible for decisions explicitly escalated to them. The system owner should remain accountable for establishing and maintaining the governance conditions under which these actors operate.

Singh et al. (2019) argue that decision provenance can support accountability by tracing how information and decisions move across interconnected systems. This principle is highly relevant to multi-agent AI because accountability depends on being able to reconstruct relationships between agents rather than simply inspecting their isolated outputs.

For every consequential decision, the accountability chain should make it possible to answer several questions

- Who initiated the workflow?
- Which agents contributed to the decision?
- What authority did each agent possess?
- Which tasks or permissions were delegated?
- What information influenced the decision?
- Which agent verified the proposed action?
- Was human approval required?
- Did any human supervisor intervene?
- Which actor executed the final action?
- Which organizational function remains answerable for the outcome?

These questions help distinguish the origin of an error from the actor that merely performed the final action.

The accountability chain should also be established before deployment rather than constructed only after an incident. Clear role allocation makes it possible to determine what each participant is expected to do, what authority it possesses, and under what conditions responsibility transfers to another actor.

Accountability Assignment Matrix

A structured accountability matrix can translate these principles into operational governance. It clarifies which actors are responsible, accountable, consulted, informed, or subject to review for different categories of multi-agent activity.

The matrix demonstrates that accountability should change according to the nature and impact of the activity. A low-risk operational action may be executed autonomously, while a high-impact decision should retain explicit human responsibility and organizational accountability. Similarly, agents may be permitted to delegate tasks only within approved limits, while system owners remain accountable for establishing and enforcing those limits.

The matrix should not be interpreted as transferring organizational accountability to artificial agents. Its purpose is to identify operational responsibility within the system while preserving a human and institutional structure of answerability. Raji et al. (2020) argue that

Table 2: Accountability Assignment Matrix for Multi-Agent AI

Governance Activity	*AI Agent*	*Human Supervisor*	*System Owner*	*Risk/Compliance Function*	*Auditor*
Routine low-risk decision	Responsible	Informed	Accountable	Monitored	Reviewable
High-risk recommendation	Responsible	Accountable	Consulted	Consulted	Reviewable
High-impact final decision	Advisory	Responsible	Accountable	Consulted	Reviewable
Agent-to-agent delegation	Responsible within approved limits	Monitored	Accountable	Policy oversight	Reviewable
Human override	Suspended or restricted	Responsible	Accountable	Informed	Reviewable
Policy configuration	Not authorized	Consulted	Responsible	Accountable	Reviewable
Audit-log review	Logged participant	Consulted	Responsible	Responsible	Accountable

effective algorithmic auditing requires responsibility to be considered across the complete organizational and technical process. Applying this principle to multi-agent AI strengthens the connection between authority, responsibility, and post-decision review.

Auditability and Decision Provenance

Minimum Audit Record

Accountability is difficult to enforce if an organization cannot reconstruct how a decision was produced. Auditability therefore forms a central part of governance in multi-agent AI. Traditional event logs may record individual system activities, but multi-agent environments require a more complete record because decisions can involve several agents, external tools, delegated permissions, policy checks, and human interventions.

Raji et al. (2020) advocate an end-to-end approach to algorithmic auditing that examines the full lifecycle and organizational context of automated decision systems. In multi-agent AI, this principle requires records that capture the complete decision process from initiation to outcome.

For consequential decisions, the minimum audit record should include

- a unique decision or workflow identifier;
- timestamps for significant actions;
- identity of the human, system, or agent initiating the workflow;
- identities and assigned roles of participating agents;
- model and agent versions;
- relevant prompts, system instructions, and governing policies;
- data sources and retrieved information;
- external tool and API interactions;
- agent-to-agent delegation events;
- permissions granted or transferred;
- intermediate recommendations;
- verification results;
- risk classifications;
- confidence or uncertainty indicators where available;
- disagreements or conflicts between agents;
- relevant policy checks;
- escalation events;
- human approvals or rejections;
- override actions;
- the final executing actor;
- the final decision or action;
- the resulting outcome.

These records should make it possible to determine not only what happened, but also whether each action was authorized and whether governance procedures were followed.

Version information is particularly important. Models, tools, system prompts, access rules, and organizational policies may change over time. Without records showing which versions were active when a decision occurred, an organization may be unable to reconstruct the conditions under which the decision was made.

Audit records should therefore capture both the decision itself and the context that shaped it.

Decision Provenance Graph

Decision provenance provides a structured representation of how information, authority, recommendations, approvals, and actions move through a multi-agent system. It is particularly valuable when a final outcome depends on several agents whose contributions cannot be understood by examining only the last output.

Singh et al. (2019) describe decision provenance as a means of tracing the movement of data and decisions across interconnected systems. In multi-agent AI, provenance should show how a request enters the system, how authority is distributed among agents, what information each agent contributes, how recommendations are verified, and how the final action is authorized.

As shown in Figure 3, the process begins with a human request or system trigger. A planner agent coordinates the workflow and assigns tasks to specialist agents.

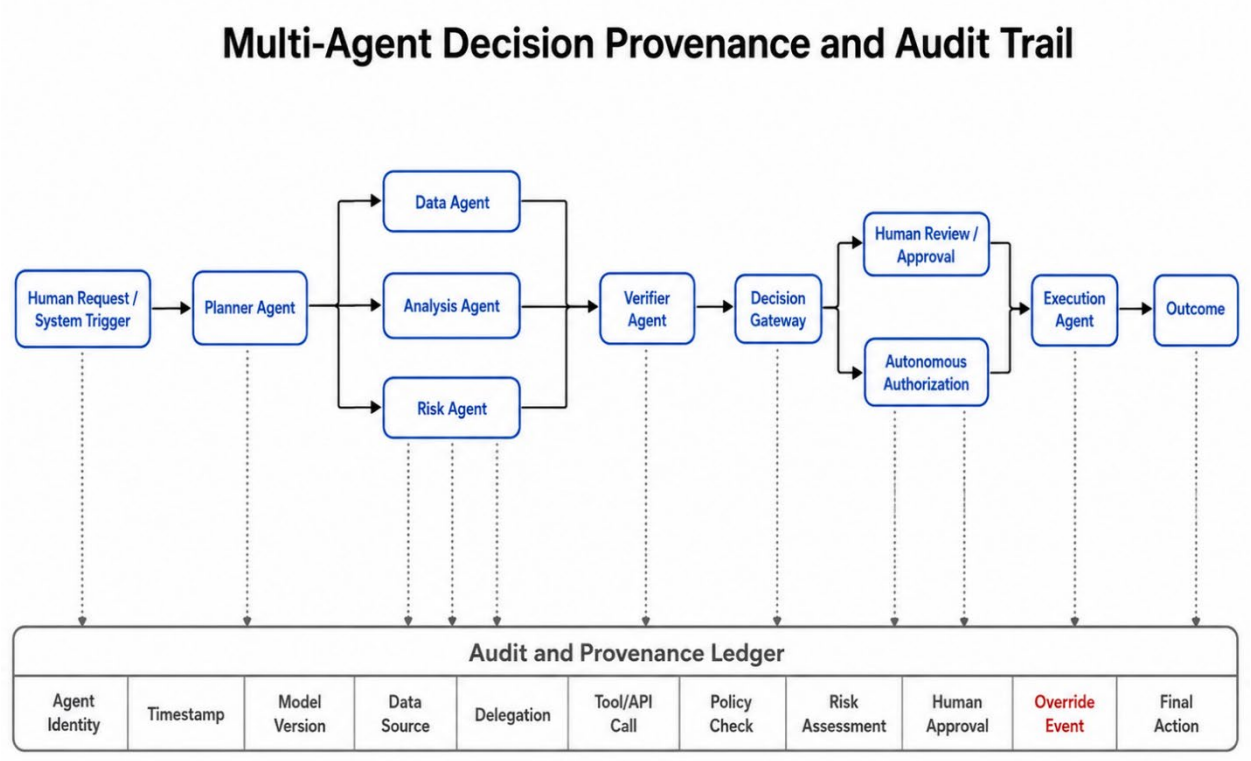

Fig 2: Multi-Agent Decision Provenance and Audit Trail showing the flow of authority, agent interactions, human review, and auditable decision records.

These specialists may include a data agent responsible for retrieving information, an analysis agent responsible for interpreting evidence, and a risk agent responsible for identifying possible harm or policy concerns. Their outputs are passed to a verifier agent, which evaluates the proposed action before it reaches the decision gateway.

At the decision gateway, the system determines whether the action may proceed under existing authority or requires human review. If authorization is granted, the execution agent carries out the action and produces the final outcome. Throughout this process, an audit and provenance ledger records the information needed to reconstruct the decision.

Chan et al. (2024) emphasize the importance of visibility into AI agents as they become increasingly capable of taking autonomous actions. In a multi-agent environment, visibility must cover not only individual agents but also their relationships. A reviewer should be able to determine which agent influenced another, how authority was delegated, which information shaped the recommendation, and whether the final action remained within established policy.

Decision provenance therefore connects technical traceability with organizational accountability. It provides evidence of who participated, what authority was exercised, and how the final decision emerged.

Auditability Principles

Effective auditability requires more than storing large volumes of system logs. Records must be accurate, interpretable, protected from unauthorized alteration, and sufficiently complete to support independent review.

The first principle is tamper resistance. Audit records should be protected against unauthorized modification so that investigators can rely on them as evidence. Appropriate controls may include restricted access, integrity verification, secure storage, and separation between operational systems and audit repositories.

The second principle is end-to-end traceability. Audit records should cover the complete decision process rather than only the final action. This includes delegation, data retrieval, external tool use, policy checks, agent disagreements, human approvals, and execution events. A correct final output does not necessarily mean that the process complied with organizational requirements.

The third principle is identity attribution. Every significant action should be associated with a specific agent, human actor, service, or technical component. Agent identities should also be linked to model and configuration versions so that later changes do not obscure the source of past actions.

The fourth principal concerns policy and configuration history. Organizations should retain the policy rules,

Table 3: Evaluation Criteria for the Proposed Governance Model

Evaluation Dimension	*Evaluation Question*	*Expected Governance Outcome*
Authority clarity	Is it clear which human or agent may make each decision?	Explicit and enforceable decision rights
Delegation control	Can agents delegate only within approved authority?	Bounded and traceable delegation
Human control	Can humans intervene before unacceptable consequences occur?	Effective override and escalation
Escalation effectiveness	Are uncertain or high-risk decisions transferred appropriately?	Risk-sensitive human involvement
Accountability	Can responsibility be assigned after an adverse outcome?	Clear responsibility and answerability
Traceability	Can the complete decision sequence be reconstructed?	End-to-end decision provenance
Auditability	Can independent reviewers verify compliance?	Reliable and protected audit evidence
Explainability	Can the organization explain how the outcome was produced?	Understandable decision reasoning and process
Adaptability	Can authority change according to context and risk?	Dynamic governance controls
Operational efficiency	Does governance preserve useful autonomy?	Proportionate oversight without unnecessary delay

permission structures, system instructions, and relevant configurations that were active when a decision occurred. This is essential because a later policy revision may otherwise make the historical decision difficult to evaluate.

The fifth principle is decision reconstruction. Where technically feasible, the audit record should allow investigators to reconstruct the important stages of a workflow. Exact reproduction may not always be possible because some AI systems produce probabilistic outputs, but sufficient evidence should remain to explain the principal steps, inputs, and authority transfers that produced the outcome.

The sixth principle is controlled audit access. Detailed logs may contain personal information, security data, commercial information, or other sensitive records. Auditability should therefore be implemented alongside appropriate privacy, confidentiality, retention, and access-control requirements.

The seventh principle is independent reviewability. Mökander et al. (2021) argue that auditing provides a structured method for assessing whether automated decision systems conform to ethical and governance requirements. Multi-agent governance should therefore enable independent auditors, compliance teams, or risk functions to review relevant records without relying solely on explanations provided by system developers or operators.

Finally, auditability should support continuous governance rather than serve only as a mechanism for investigating failures. Provenance records can help organizations identify repeated authority violations, unusual delegation patterns, recurring agent disagreements, weaknesses in verification processes, and failures to escalate high-risk decisions. These observations can then inform changes to decision rights, policies, agent permissions, and supervisory arrangements.

Governance Model Evaluation

The proposed governance model should be evaluated according to its ability to control decision authority, preserve human accountability, and maintain reliable evidence of how multi-agent decisions are produced. Because the framework is conceptual, evaluation should focus on whether its governance mechanisms address the main risks identified in the literature and whether they can be applied consistently across different organizational settings.

The first evaluation dimension is authority clarity. Every agent and human participant should have a defined level of decision authority. The model should make clear which actors may recommend, approve, execute, delegate, monitor, or override particular actions. This is necessary because unclear authority can create situations in which several agents contribute to a decision without any actor having clear responsibility for the final outcome. The distinction between formal and real authority identified by Aghion and Tirole (1997) is important here, since organizations must ensure that formally assigned decision rights reflect the way authority is actually exercised in practice.

The second dimension is delegation control. Multi-agent systems should permit delegation only within predefined boundaries. An agent should not be able to transfer authority that it does not itself possess. The framework should therefore be capable of identifying unauthorized delegation, excessive sub-delegation, and the use of tools or data outside approved limits.

The third dimension is human control. The model should allow authorized human actors to intervene before unacceptable consequences become irreversible. Santoni de Sio and van den Hoven (2018) argue that meaningful human control depends on a real connection between human intentions, system behavior, and responsibility. Evaluation should therefore examine whether supervisors receive enough information, whether escalation occurs early enough, and whether humans possess practical authority to pause, reject, restrict, or terminate actions.

The fourth-dimension concerns accountability and traceability. It should be possible to reconstruct who participated in a decision, what authority each participant possessed, what information was used, and how the final action was approved. Singh et al. (2019) show the importance of decision provenance for tracing complex decision processes, while Raji et al. (2020) emphasize end-to-end auditing rather than isolated examination of final model outputs.

Finally, the framework should be evaluated for operational usefulness. Governance controls should not make routine multi-agent activity unnecessarily slow or dependent on constant human approval. A successful model should permit low-risk autonomy while increasing oversight as uncertainty, impact, or policy sensitivity grows.

The model can also be assessed using scenario-based evaluation. Representative cases may include financial transactions, cybersecurity incident response, automated recruitment, supply-chain procurement, or autonomous software deployment. Each scenario can be tested to determine whether the framework correctly identifies the permitted level of autonomy, required human involvement, accountable actor, and necessary audit evidence.

Discussion

The findings developed throughout this study indicate that the governance of multi-agent AI cannot be reduced to a simple choice between human control and machine autonomy. The more important issue is how authority should be distributed across different actors and how that authority should change when risk, uncertainty, or potential harm increases.

One of the central tensions in multi-agent AI is the balance between autonomy and control. Greater autonomy can improve speed, scalability, and operational efficiency because agents can coordinate and execute tasks without waiting for repeated human approval. However, increased autonomy can also reduce visibility and create uncertainty about responsibility when several agents contribute to the same outcome. Chan et al. (2023) identify this broader problem in increasingly agentic systems, where the ability to pursue goals and act with reduced supervision can create new forms of harm.

The proposed governance model addresses this tension by treating autonomy as a bounded and conditional form of authority rather than an unrestricted technical capability. Agents may operate independently where the potential impact is low and decision boundaries are clearly defined. As risk rises, the model increases verification, escalation, and human involvement. High-impact or irreversible decisions can therefore remain under direct human authority.

This approach also challenges the assumption that meaningful oversight requires continuous human participation. Parasuraman et al. (2000) show that different levels of automation can be appropriate for different functions. Applied to multi-agent AI, this means that humans do not need to approve every routine action. Instead, governance should determine where human judgment is most necessary and ensure that supervisors are given both the information and authority required to intervene effectively.

A second important implication concerns accountability. Multi-agent systems distribute operational activity, but they should not distribute accountability to the point where no identifiable actor remains answerable. Novelli et al. (2024) emphasize that accountability requires identifiable relationships of responsibility and answerability. The proposed model therefore maintains organizational accountability even when individual agents perform substantial portions of a workflow.

The study also shows that auditability is not simply an administrative requirement. It is part of the governance architecture itself. Decision provenance allows organizations to examine whether authority was used properly, whether delegation remained within approved limits, whether human approval was required, and whether agents complied with applicable policies. Without these records, accountability may become largely symbolic because the organization may be unable to establish what actually occurred.

From a practical perspective, the framework provides guidance for chief AI officers, risk managers, compliance teams, system architects, internal auditors, and senior management. Organizations deploying multi-agent systems should establish explicit authority boundaries before deployment, define escalation thresholds, assign accountable owners, and maintain records capable of reconstructing consequential decisions.

The framework also has implications for regulators and policymakers. Requirements for human oversight should consider whether humans possess meaningful intervention powers rather than merely appearing somewhere in the workflow. Similarly, auditability

requirements should examine the complete decision chain rather than focusing only on model explanations or final outputs.

Challenges and Limitations

The proposed governance model faces several limitations that should be considered when applying it in practice.

One challenge is the difficulty of defining appropriate decision boundaries. Risk varies across industries, organizations, and operational contexts. An action considered routine in one environment may have serious legal, financial, or safety consequences in another. The framework therefore requires organizational adaptation rather than a single universal threshold.

A second limitation concerns the complexity of agent interactions. Large multi-agent systems may generate substantial volumes of messages, tool calls, intermediate decisions, and delegation events. Recording every interaction can create storage, privacy, and analysis challenges. Excessive logging may also reduce the usefulness of audit records if important events become difficult to identify.

Human oversight introduces additional difficulties. Supervisors may experience information overload, automation bias, or excessive reliance on agent recommendations. Green (2022) argues that human oversight can be ineffective when humans lack sufficient information or practical ability to challenge automated outcomes. This means that the presence of a human approval stage alone does not guarantee effective governance.

Another limitation concerns timing. Some multi-agent applications may operate at speeds where human approval creates significant delays. Cybersecurity response, financial trading, or real-time infrastructure management may require decisions within seconds or milliseconds. Governance mechanisms must therefore distinguish between actions that genuinely require human approval and those that can be controlled through predefined restrictions and post-action review.

Technical limitations also remain. Proprietary models may provide limited access to internal reasoning or system configuration. External tools and third-party services can create additional gaps in provenance. If an agent invokes an external system that does not maintain comparable audit records, the decision chain may become incomplete.

Privacy is another concern. Detailed provenance records can contain personal data, confidential business information, security details, or sensitive system instructions. Organizations must balance auditability with data minimization, retention requirements, and access controls.

The conceptual nature of this study is also a limitation. The proposed framework has not yet been tested across large-scale operational deployments. Further empirical evaluation is therefore needed to determine how well the model performs under real organizational conditions.

Future Research Directions

Future research should extend the proposed governance model through empirical, technical, and organizational studies.

First, the framework should be validated using real or simulated multi-agent systems in sectors such as finance, healthcare, cybersecurity, recruitment, public administration, and supply-chain management. Such studies could examine whether the proposed authority levels and escalation mechanisms reduce governance failures without creating excessive operational delays.

Second, future work should develop quantitative measures of agent autonomy and decision risk. Current governance approaches often rely on broad categories such as low, medium, and high risk. More precise measures could help organizations adjust decision rights dynamically according to uncertainty, potential impact, reversibility, and policy sensitivity.

Third, researchers should examine machine-readable governance policies. Decision rights, escalation thresholds, and delegation constraints could potentially be expressed in formats that agents and governance systems can enforce automatically.

Fourth, further research is needed on **agent identity and authorization infrastructure**. Multi-agent environments require reliable methods for verifying agent identity, determining permissions, and recording how authority is transferred between agents.

Fifth, cryptographically verifiable provenance could strengthen audit integrity. Techniques such as signed event records, secure timestamps, and tamper-evident logs may improve confidence that decision histories have not been altered after an incident.

Sixth, future studies should investigate governance mechanisms for self-modifying or dynamically generated agents. These systems may create new challenges because their behavior, capabilities, or identities can change during operation.

Seventh, cross-organizational multi-agent ecosystems require greater attention. Agents operated by different companies may collaborate within the same workflow, creating difficult questions about jurisdiction, responsibility, data access, and audit rights.

Eighth, human factors research should examine how supervisors respond to agent escalation. Issues such as alert fatigue, automation bias, trust, expertise, and workload may strongly affect whether human override functions as intended.

Ninth, further work should examine governance of very large agent populations, including agent swarms and decentralized coordination structures. Traditional

role-based accountability may become difficult when hundreds or thousands of agents contribute to an outcome.

Finally, future research should support the development of standardized multi-agent audit protocols. Common requirements for decision provenance, delegation records, human intervention logs, and authority verification would help organizations assess multi-agent systems more consistently.

Conclusion

Multi-agent artificial intelligence introduces a governance challenge that is fundamentally different from the management of conventional automated systems. Decisions may no longer be produced by a single model or clearly identifiable actor. Instead, they may emerge from interactions among planners, specialist agents, verification agents, execution agents, external tools, and human supervisors. This distributed structure can improve operational capability, but it also creates uncertainty about authority, responsibility, intervention, and accountability.

This study proposed a governance model centered on four connected requirements: decision rights, human override, risk-based escalation, and end-to-end auditability. The model treats autonomy as a controlled form of authority that should be assigned according to decision type, risk, uncertainty, reversibility, and organizational policy. It also establishes that delegation between agents must remain bounded and traceable.

Human involvement is treated as a proportionate governance mechanism rather than a universal approval requirement. Low-risk activities may proceed autonomously, while high-impact, uncertain, or irreversible actions should trigger greater human involvement. Effective oversight therefore depends on giving supervisors sufficient information and practical authority to intervene when necessary.

The study further argues that accountability should follow the complete decision chain. Operational responsibility may be distributed among agents, but organizational accountability should remain identifiable. Decision provenance and protected audit records provide the evidence needed to reconstruct how authority was exercised and how outcomes were produced.

The central governance question for multi-agent AI is therefore not simply whether artificial agents should be allowed to act autonomously. It is which actor has the right to make a particular decision, under what conditions, with what authority, subject to what form of human intervention, and with what evidence available for later review.

By connecting these elements within a single governance structure, the proposed model provides a foundation for organizations seeking to use multi-agent AI while preserving human authority, institutional accountability, and reliable oversight.

References

Aghion, P., & Tirole, J. (1997). Formal and real authority in organizations. *Journal of Political Economy, 105*(1), 1–29. DOI: 10.1086/262063.

Jadala, S. K. (2026). Agentic Artificial Intelligence in Cybersecurity: Emerging Risks, Governance Challenges, and Defensive Frameworks. International Journal of Artificial Intelligence and Engineering Research, 2(02).

Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., Langosco, L., He, Z., Duan, Y., Carroll, M., Lin, M., Mayhew, A., Collins, K., Molamohammadi, M., Burden, J., Zhao, W., Rismani, S., Voudouris, K., Bhatt, U., … Maharaj, T. (2023). Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 651–666). Association for Computing Machinery. DOI: 10.1145/3593013.3594033.

Chan, A., Ezell, C., Kaufmann, M., Wei, K., Hammond, L., Bradley, H., Bluemke, E., Rajkumar, N., Krueger, D., Kolt, N., Heim, L., & Anderljung, M. (2024). Visibility into AI agents. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 958–973). Association for Computing Machinery. DOI: 10.1145/3630106.3658948.

Potla, R. B. (2024). Optimizing extended warehouse management for make-to-order plants: Slotting, wave picking, and yard orchestration at scale. Journal of Computer Science and Technology Studies, 6(3), 181-192.

Cobbe, J., Lee, M. S. A., & Singh, J. (2021). Reviewable automated decision-making: A framework for accountable algorithmic systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 598–609). Association for Computing Machinery. DOI: 10.1145/3442188.3445921.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM, 59*(2), 56–62. DOI: 10.1145/2844110.

Kunaparaju, C. (2025). AI-Driven Cyber Defense Systems: Strengthening National Security through Intelligent Threat Prediction and Response. Algora, 2(1), 1-30.

Begimher, D., Leo, C., Huang, J., Gaw, P., & Zheng, B. (2026). SIR-Bench: Evaluating Investigation Depth in Security Incident Response Agents. arXiv preprint arXiv:2604.12040.

Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., López de Prado, M., Herrera-Viedma, E., & Herrera, F. (2023). Connecting the dots in trustworthy artificial intelligence: From AI principles, ethics, and key requirements to responsible AI systems

and regulation. *Information Fusion, 99*, 101896. DOI: 10.1016/j.inffus.2023.101896.

Fama, E. F., & Jensen, M. C. (1983). Separation of ownership and control. *The Journal of Law and Economics, 26*(2), 301–325. DOI: 10.1086/467037.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. DOI: 10.1007/s11023-018-9482-5.

Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review, 45*, 105681. DOI: 10.1016/j.clsr.2022.105681.

Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., Smith, C., Barfuss, W., Foerster, J., Gavenčiak, T., Han, T. A., Hughes, E., Kovařík, V., Kulveit, J., Leibo, J. Z., Oesterheld, C., Schroeder de Witt, C., Shah, N., Wellman, M., … Rahwan, I. (2025). *Multi-agent risks from advanced AI* (Technical Report No. 1). Cooperative AI Foundation. DOI: 10.48550/arXiv.2502.14143.

Jennings, N. R. (2000). On agent-based software engineering. *Artificial Intelligence, 117*(2), 277–296. DOI: 10.1016/S0004-3702(99)00107-1.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. DOI: 10.1038/s42256-019-0088-2.

Kolt, N. (2025). Governing AI agents. *Notre Dame Law Review, 101*, forthcoming. DOI: 10.48550/arXiv.2501.07913.

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review, 165*(3), 633–705.

Leo, C., Dykyi, A., Cortegaca, D., Begimher, D., & Jha, P. (2026). ThreatForest: Multi-Agent Attack Tree Generation with Pluggable TTP Framework Mapping. arXiv preprint arXiv:2607.27528.

Laux, J. (2024). Institutionalised distrust and human oversight of artificial intelligence: Towards a democratic design of AI governance under the European Union AI Act. *AI & Society, 39*, 2853–2866. DOI: 10.1007/s00146-023-01777-z.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence, 1*(11), 501–507. DOI: 10.1038/s42256-019-0114-4.

Mökander, J., Morley, J., Taddeo, M., & Floridi, L. (2021). Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations. *Science and Engineering Ethics, 27*(4), Article 44. DOI: 10.1007/s11948-021-00319-4.

Novelli, C., Taddeo, M., & Floridi, L. (2024). Accountability in artificial intelligence: What it is and how it works. *AI & Society, 39*, 1871–1882. DOI: 10.1007/s00146-023-01635-y.

Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems, 34*(2), 101885. DOI: 10.1016/j.jsis.2024.101885.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, 30*(3), 286–297. DOI: 10.1109/3468.844354.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33–44). Association for Computing Machinery. DOI: 10.1145/3351095.3372873.

Sambamurthy, V., & Zmud, R. W. (1999). Arrangements for information technology governance: A theory of multiple contingencies. *MIS Quarterly, 23*(2), 261–290. DOI: 10.2307/249754.

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI, 5*, Article 15. DOI: 10.3389/frobt.2018.00015.

Potla, R. B. (2024). A SOX/ITAR-aligned global ERP template for multi-plant manufacturers: Governance patterns and controls. J Artif Intell Mach Learn & Data Sci, 2(2), 3222-3232.

Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., Lee, T., Mishkin, P., Eloundou, T., Hickey, A., Slama, K., Ahmad, L., McMillan, P., Beutel, A., Passos, A., & Robinson, D. G. (2023). *Practices for governing agentic AI systems*. OpenAI.

Singh, J., Cobbe, J., & Norval, C. (2019). Decision provenance: Harnessing data flow for accountable systems. *IEEE Access, 7*, 6562–6574. DOI: 10.1109/ACCESS.2018.2887201.

Sterz, S., Baum, K., Biewer, S., Hermanns, H., Lauber-Rönsberg, A., Meinel, P., & Langer, M. (2024). On the quest for effectiveness in human oversight: Interdisciplinary perspectives. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2495–2507). Association for Computing Machinery. DOI: 10.1145/3630106.3659051.

Verdiesen, I., Santoni de Sio, F., & Dignum, V. (2021). Accountability and control over autonomous weapon systems: A framework for comprehensive human oversight. *Minds and Machines, 31*(1), 137–163. DOI: 10.1007/s11023-020-09532-9.

Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 1–18). Association for Computing Machinery. DOI: 10.1145/3351095.3372833.

Winfield, A. F. T., & Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2133), 20180085. DOI: 10.1098/rsta.2018.0085.

Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *The Knowledge Engineering Review, 10*(2), 115–152. DOI: 10.1017/S0269888900008122.

Potla, R. (2023). Designing a BTP-centric integration mesh for shop-floor IoT, MES and ERP in discrete manufacturing. Journal of Artificial Intelligence, Machine Learning and Data Science, 1(2), 1-8.