From Agents to Viable Collectives: The VSM as a Diagnostic Framework

Krishan Mathis; Margeret Heath; Panos Panagiotakopoulos

Draft Paper — not peer reviewed

2026-08-24

Contents

Abstract

Research on artificial intelligence agents commonly evaluates the capabilities, incentives, and interactions of individual agents. Yet increasingly consequential systems are not individual agents but collectives: ensembles of language-model agents, hybrid human–AI teams, and organizational arrangements in which observations, recommendations, decisions, and actions are distributed across people and synthetic components. Strong constituent agents do not necessarily produce a coherent, adaptive, or accountable collective. This article develops the Viable System Model (VSM) as a diagnostic framework for that organizational problem. It distinguishes AI agents, agent-based models, and deployed agentic systems; defines collec-

tive viability as the continued maintenance of a distinct organization under changing conditions; and translates the VSM's functions of operation, coordination, regulation, intelligence, and identity into research problems for hybrid human–synthetic systems. The proposed mapping is functional rather than ontological: it neither assumes that every agent collective instantiates the VSM nor assigns one agent to each VSM function. The synthesis identifies six hard problems: recursive boundary formation, collective identity, channel capacity, conflict mediation, exception handling, and distributed accountability. From these it formulates falsifiable propositions and demonstrator designs. The resulting agenda shifts evaluation from isolated agent capability toward the organizational conditions under which populations of agents can remain effective, adaptive, governable, and answerable.

Keywords: viable system model; AI agents; multi-agent systems; hybrid intelligence; organizational cybernetics; collective intelligence; AI governance; sociotechnical systems

---

1. Introduction: from capable agents to viable collectives

Artificial intelligence is increasingly organized as populations rather than deployed as a single model. Language-model agents are assigned specialized roles, equipped with tools and memories, and placed in conversational or procedural relationships with other agents. Humans may specify goals, approve exceptional actions, supply contextual judgment, or participate as members of the same operational team. The relevant object of analysis is therefore often not an artificial agent in isolation but a hybrid human–synthetic collective: a bounded arrangement of people, artificial agents, technical infrastructures, and institutional rules that jointly produces consequential action.

This change of scale creates a problem that individual-agent evaluation cannot resolve. An agent can be competent at planning, reasoning, or tool use while the larger arrangement duplicates work, suppresses dissent, loses information, terminates prematurely, or cannot determine who may legitimately alter its goals. Recent results make this distinction empirically salient. Cemri et al. (2025) identified fourteen failure modes across 1,642 execution traces from seven multi-agent language-model systems. The failures clustered around system specification, inter-agent misalignment, and task verification and termination; changing the system architecture altered the distribution of failures even when the underlying model was held constant. Zomer and De Domenico (2026) likewise found that stronger individual

decision capabilities did not automatically yield better collective performance: the consensus tendencies of language-model agents could produce premature convergence, and changes to communication topology altered the trade-off between exploration and convergence. In human–AI teams, even perceived agent identity can change communication and performance despite unchanged teammate competence (Qin, Lee, and Sajda 2026). These findings point to an organizational, not merely cognitive, research problem.

The distinction is important because contemporary research often asks whether an agent can complete a task, whether several agents outperform one, or which interaction protocol raises a benchmark score. These are necessary questions, but they do not establish whether a collective can preserve coherent operations, adapt to novel conditions, maintain legitimate purpose, or account for its actions across time. Short-horizon task success is not the same as viability. Nor does emergence guarantee an organization capable of governing its own emergence.

Organizational cybernetics offers a vocabulary for investigating this missing level of analysis. Stafford Beer's Viable System Model describes a set of interdependent functions that a system must perform if it is to maintain a separate existence in a changing environment (Beer 1979, 1981, 1984, 1985). In simplified form, these concern the autonomy of primary operations, coordination among those operations, internal regulation and assurance, intelligence about the environment and future, and identity or policy. The model is recursive: viable operational units are themselves systems embedded in larger viable systems. It also treats communication capacity and the management of variety as constitutive of organization rather than as implementation details.

This article argues that the VSM can be used as a functional diagnostic framework for hybrid human–synthetic collectives. Its core claim is that collective viability cannot be inferred from the capabilities of constituent agents. It depends on whether the collective performs a set of interdependent organizational functions under changing conditions. The VSM does not supply a ready-made software architecture, and this article does not claim that each function should be assigned to a separate artificial agent. Instead, the model is used to ask what must be achieved somewhere in the organization, how those achievements depend on one another, and how their absence could be observed.

This argument makes three contributions. First, it separates AI agents, agent-based models, and deployed agentic systems, which differ in purpose, evidential status, and locus of accountability. Second, it translates VSM functions into open problems and observable failure conditions without asserting a one-to-one equivalence. Third, it formulates a falsifiable re-

search programme for comparing alternative organizational architectures in simulations and field demonstrators.

The novelty claim is deliberately bounded. The VSM has been applied to agent systems before, in holonic and multi-agent manufacturing (Herrera, Belmokhtar Berraf, and Thomas 2012), multi-robot adaptation (Ardavs et al. 2019), and AI-augmented organizations (Pérez Ríos 2025; Kuhr, Mennenga, and Herrmann 2026). Three things distinguish the present article from that work. It addresses language-model agents rather than manufacturing holons or robot teams. It asks about legitimacy and accountability alongside viability, rather than treating organizational coherence as sufficient. And it states in advance what evidence would show the diagnosis to be wrong.

2. Three objects too often conflated

The terms *agent*, *agent-based model*, and *agentic system* refer to different objects. Their conflation encourages invalid transfers of evidence and obscures where organizational responsibility lies.

2.1 AI agents

An AI agent is generally characterized by its capacity to receive information about an environment, select actions, and pursue objectives with some degree of autonomy. Classical accounts emphasize autonomy, social ability, reactivity, and proactiveness (Wooldridge and Jennings 1995). Contemporary language-model agents extend this formulation through natural-language planning, memory, tool use, and interaction with other agents. These capabilities vary in implementation and degree; "agent" is therefore a functional designation, not a claim about consciousness, personhood, or moral status.

Agent-level evaluation typically examines task completion, reasoning accuracy, planning efficiency, tool-use reliability, or compliance with instructions. Such measures are appropriate for an individual component. They become insufficient when several agents share information, depend on one another's outputs, or alter a common environment. The collective may exhibit properties that cannot be located in any single agent: oscillation, conformity, cascading error, bottlenecks, institutional drift.

2.2 Agent-based models

An agent-based model (ABM) is a scientific representation in which heterogeneous entities follow specified rules and interact in a simulated environment so that macro-level patterns can be studied (Macal and North 2010). Its agents may be intentionally simple. The purpose is not necessarily to

build an operational autonomous system but to investigate how assumptions about micro-level behavior generate system-level outcomes.

The epistemic status of an ABM therefore differs from that of a deployed agentic arrangement. A simulation can test whether a proposed mechanism is sufficient to generate an observed pattern, compare counterfactual structures, or expose sensitivity to assumptions. It does not by itself establish that the represented entities possess the modeled capacities or that a deployed organization will behave equivalently. Generative agents powered by language models blur this boundary because they can be used both as simulation subjects and as operational components. The distinction must nevertheless be retained: an agent used to represent a person in a model is not the same object as an agent authorized to act in a public-service workflow.

2.3 Agentic systems

An agentic system is an operational arrangement in which one or more artificial agents participate in a continuing process of observation, decision, communication, or action. Its boundary may also include human operators, managers, affected communities, databases, tools, legal mandates, and escalation procedures. The system's behavior is produced through relations among these elements rather than by an agent alone.

The vocabulary follows Chan et al. (2023), who fix the grammar (agency is the property, agent the role, agentic the adjective) and who decline to treat agency as a binary attribute, identifying instead four characteristics that admit of degree: underspecification, directness of impact, goal-directedness, and long-term planning. Shavit et al. (2023) define agenticness compatibly, as the degree to which a system can adaptably achieve complex goals in complex environments with limited direct supervision.

The adjective is borrowed, and its earlier senses do not all point the same way. In personality and social psychology it descends from Bakan (1966), and Bandura (2001) established the sense the AI literature inherited, agency as intentional and self-directed influence, in an argument built expressly against computational metaphors of mind. Nothing in the present use claims that sense for software. The usage here is closer to Emirbayer and Mische (1998), for whom agency is not a substance an actor possesses but a temporally structured process of engagement, and to Rammert (2008), who grades agency and locates it in hybrid constellations of people, machines, and programs rather than in any single component.

This definition makes two analytical moves. First, it shifts attention from component properties to organizational relations. The same model may behave differently when placed in a sequential workflow, a peer network,

a hierarchy, or a market. This is not a speculative point. Instances drawn from one underlying system can be given different tools and conditioned by different prompts, so that agenticness is a property of the deployment rather than of the model (Chan et al. 2024). Second, it treats the surrounding institutional arrangement as part of the system under study. A human approval checkpoint, an audit trail, a procurement rule, or a statutory duty is not external "context" when it causally shapes what the collective can do.

The relevant unit of analysis must therefore be declared rather than assumed. For some questions it is the artificial agent; for others, the conversational ensemble; for still others, the human organization in which the ensemble operates. This requirement anticipates the VSM concept of recursion: an entity may be operationally autonomous at one level while remaining a component of a broader organization at another.

3. Collective viability as the missing analytical problem

Beer defined viability in terms of maintaining a separate existence (Beer 1984). This definition is more demanding than remaining active and less normatively flattering than the ordinary language of "healthy" organizations. A harmful organization may be viable. Conversely, an ethically legitimate project may be non-viable if it cannot coordinate operations, learn, or secure the resources needed to continue. Viability is therefore a descriptive property that must be evaluated alongside, not substituted for, legitimacy.

Several neighboring concepts should be distinguished. *Persistence* denotes continuation through time but says little about the mechanisms or identity involved: a dysfunctional process can persist through inertia. *Robustness* is the capacity to preserve performance under a specified range of perturbations without fundamental reorganization. *Resilience* often emphasizes recovery or transformation after disturbance. *Optimization* improves performance against an objective function, potentially at the cost of unmeasured capabilities or future adaptability. Viability concerns whether an organization can continue as a recognizable entity while regulating present operations and adapting its relationship with a changing environment. Robustness and resilience may support viability, but neither is identical to it.

This distinction exposes the limits of common agent benchmarks. A collective can maximize a task score while consuming unsustainable resources, narrowing its information sources, or losing the ability to respond to novelty. It may appear robust against anticipated faults while failing when the definition of the task changes. It can coordinate efficiently around an illegitimate or obsolete objective. Evaluating viability therefore requires multiple timescales and a theory of identity: what must remain invariant for the

collective to count as the same organization, and what must be allowed to change?

The problem is compounded in hybrid collectives. Artificial agents may be replaceable, replicated, or reassigned. Humans participate through roles that carry different kinds of authority. Organizational boundaries can cut across firms and platforms; one agent may serve several collectives simultaneously. Goals can be embedded in prompts, policies, reward functions, contracts, and professional norms, none of which alone constitutes legitimate purpose. These conditions make collective identity an accomplishment of governance rather than an emergent label.

The VSM is relevant because it treats viability as a relationship among differentiated functions. Operational autonomy without coordination can fragment the organization. Coordination without autonomy can suppress the local variety needed for effective action. Internal regulation without environmental intelligence can optimize an obsolete present. Intelligence without identity can generate alternatives without a legitimate basis for choosing among them. Identity without operational feedback becomes symbolic policy detached from what the system actually does.

This functional interdependence suggests a different research question. Instead of asking whether multiple agents are collectively intelligent in the aggregate, we ask whether the organization can repeatedly produce coherent action, detect and absorb relevant disturbances, reconsider its models, preserve or legitimately revise its identity, and account for exceptional decisions. Collective viability is not a scalar capability added to the scores of individual agents but a pattern of organization that has to be sustained through time.

4. The VSM as a functional diagnostic framework

The VSM is frequently depicted as five numbered systems. Read too literally, this diagram invites a superficial exercise in which existing roles or software components are relabeled System 1 through System 5. The more useful interpretation begins with functions and relationships. The question is not which component resembles a VSM box, but whether the organization performs the necessary regulatory work and whether the channels among those functions have sufficient capacity.

4.1 System 1: primary operations and local autonomy

System 1 comprises the primary activities through which the system-in-focus enacts its purpose. In a hybrid collective, these may be case-handling teams, scientific work packages, response units, or service-delivery processes. An

individual AI agent is not automatically a System 1 unit. The relevant unit must produce a transformation that matters to the identity of the larger system and possess enough autonomy to regulate its local work.

The design problem is autonomy within cohesion. Local units require freedom to respond to conditions that central governance cannot fully anticipate. Yet their actions create dependencies and externalities for other units. Agentic architectures often treat autonomy as a property of components. Agent theory itself has long resisted that reading: autonomy is relational, held with respect to particular decisions and particular authorities, and constrained by dependence on others (Castelfranchi 1995). The VSM reframes it as a negotiated organizational relationship. Autonomy is viable only when local variety is preserved while consequences that affect the larger system remain governable.

The word *agentic* carries this problem in its origin. Milgram (1974) introduced the agentic state to name the condition of a person who sees himself as an instrument for carrying out another's wishes, and defined it in explicit opposition to autonomy. He reached it through a cybernetic argument about self-directed automata, whose local control must be suppressed in favor of regulation by a higher-level component once they are placed in a hierarchy. What Milgram treated as a moral problem for human subordinates is the design problem stated here in another vocabulary.

Candidate observations include local task success, the proportion of decisions resolved without higher intervention, cross-unit dependencies, and the diversity of locally generated responses. Characteristic failures include brittle centralization, in which every exception waits for a bottleneck, and incoherent local optimization, in which individually successful units damage the collective outcome. Organization theory states the same tension as a joint requirement: units facing different environments must be allowed to differ, and the organization must then invest in mechanisms that reconcile them (Lawrence and Lorsch 1967). Sociotechnical systems research reached a parallel conclusion, that social and technical organization must be optimized together rather than in sequence (Trist and Bamforth 1951).

4.2 System 2: coordination, mediation, and oscillation control

System 2 dampens destructive oscillation among primary units. Coordination is broader than message exchange or task allocation. It includes shared schedules, protocols, standards, mutual adjustment, and mechanisms that mediate incompatible demands before they require authoritative intervention. In multi-agent language-model systems, relevant phenomena include repeated conversational loops, duplicate work, inconsistent state, premature consensus, and agents acting on incompatible assumptions.

The VSM perspective cautions against treating more communication as an unqualified good. Unstructured communication can increase noise, conformity, and coordination cost. The caution is supported well outside agent research. Communication topology has been known since Bavelas (1950) to shape group performance and leadership emergence independently of member ability, and even mild social influence narrows the spread of individual estimates without improving their accuracy, yielding a consensus that is more confident but no more correct (Lorenz et al. 2011). Zomer and De Domenico's (2026) findings show that network topology changes exploration–exploitation dynamics; Cemri et al. (2025) show that ignoring or withholding information can have different organizational causes despite similar surface symptoms. System 2 design therefore concerns the form, timing, and semantics of interaction, as well as who can interrupt or reframe it.

Possible measures include conflict duration, duplicate-work rates, message churn, unresolved dependencies, convergence time, and the diversity retained at convergence. Failure may appear as deadlock, thrashing, fragmentation, or an apparently efficient consensus reached before relevant alternatives have been explored.

4.3 Systems 3 and 3*: internal regulation, resource bargaining, and audit

System 3 concerns the internal coherence of the operational whole. It negotiates resources, establishes constraints, monitors aggregate performance, and intervenes when the behavior of primary units threatens collective viability. System 3* denotes supplementary audit or monitoring that can examine operational reality without relying exclusively on routine reports.

For agentic systems, this distinction is especially important. Agents may produce fluent status reports that are incorrect, incomplete, or mutually reinforcing. A verifier that merely consumes the same conversation may share its blind spots. This is consistent with evidence that language models do not reliably improve their own reasoning when revising without external feedback (Huang et al. 2024). Independent access to logs, tools, environmental state, or sampled outputs can provide a functional analogue of audit. This does not require a permanently centralized controller. Regulatory authority can be distributed, but the collective must still be able to allocate scarce resources, enforce constraints, and verify claims about its own state.

Measures may include resource-allocation latency, intervention rates, audit coverage, discrepancy detection, verification accuracy, and the proportion of operational claims supported by independent evidence. Pathologies include uncorrected drift, fabricated status, a regulator overwhelmed by low-level decisions, or auditing so intrusive that it destroys operational autonomy.

The literature on institutional audit adds a further pathology: verification detached from the work it inspects can become a ritual that certifies the existence of auditable systems rather than the quality of what they produce (Power 1997).

4.4 System 4: environmental intelligence and adaptation

System 4 models the environment and the future. It identifies changes that may make current operations or even current identity untenable, generates alternatives, and places the organization's internal model into productive tension with what the environment is becoming. Reducing this function to sensing or data retrieval misses its role in imagination, learning, and strategic redesign. Organization theory draws the same line between scanning and interpretation: what an organization notices depends on how analyzable it assumes its environment to be and how actively it intrudes into it (Daft and Weick 1984).

Many agentic systems are optimized for reactive task execution. Web search, retrieval, or monitoring can expand what they observe without ensuring that observations alter the organization's assumptions. A viable intelligence function must distinguish noise from structurally significant change, maintain alternative hypotheses, and create protected capacity for exploration. Organizational learning research explains why that capacity must be protected rather than assumed: the returns to exploitation are nearer, surer, and easier to measure, so adaptive systems drift toward it at the cost of long-run viability (March 1991). It must also communicate with present-focused regulation: adaptation that ignores operational constraints becomes fantasy, while regulation that dominates intelligence produces strategic blindness.

Candidate observations include novelty-detection rates, scenario diversity, adaptation latency, exploration effort, model revision after disconfirming evidence, and changes to operations traceable to environmental learning. Failures include stale world models, perpetual reaction, premature exploitation, and strategic analysis that never affects action.

4.5 System 5: identity, policy, and normative closure

System 5 concerns identity and policy. It balances present operations with future adaptation and establishes the values and purposes within which lower-level choices make sense. In hybrid collectives this is the point at which a functional account of viability meets political and ethical questions.

Synthetic agents can represent, compare, and apply policies. These capabilities do not by themselves authorize the policies. A system's purpose may emerge descriptively from its behavior. The purpose of a system is what it

does, in Beer's dictum (Beer 2002). Legitimate purpose is a different matter. It requires an account of who is entitled to bind whom, how affected parties can contest decisions, and how policy can be revised. Legitimacy in this sense is a generalized perception that an entity's actions are proper within a socially constructed order, and it is analytically distinct from effectiveness (Suchman 1995). Institutional analysis supplies a vocabulary for stating such constraints precisely, distinguishing shared strategies, norms, and rules by their deontic force and their sanctions (Crawford and Ostrom 1995). System 5 must therefore not be anthropomorphized as an artificial "collective self." It is better understood as the institutional processes through which identity and ultimate constraints are constituted and maintained.

Observations could include goal consistency across levels, the provenance of policy changes, participation in authorization, frequency and grounds of overrides, and the treatment of contested decisions. Failure conditions include goal drift, symbolic policy contradicted by operations, illegitimate closure around agent-generated objectives, and paralysis when identity commitments conflict.

4.6 Recursion: viability at more than one scale

Beer's recursive system theorem holds that viable systems contain and are contained within viable systems (Beer 1979; see Espejo and Reyes 2011 for a contemporary treatment). Recursion is not equivalent to a management hierarchy or a nested software call graph. It identifies repeated levels at which a unit conducts primary operations while also requiring coordination, regulation, intelligence, and identity.

Hybrid collectives make recursion difficult because membership and jurisdiction overlap. A clinical AI agent may participate in a patient-care team, a hospital service, a vendor platform, and a national regulatory regime. Each level observes different environments and bears different responsibilities. The central analytical task is to identify the system-in-focus, its primary transformations, the viable units it contains, and the wider systems that constrain it. Political economy describes a comparable arrangement as polycentric governance: multiple semi-autonomous decision centres, each with its own rules, operating within a larger order rather than under a single hierarchy (Ostrom 2010).

Measures include the clarity of jurisdiction, rates of cross-level escalation, contradictory directives, boundary changes, and the proportion of disturbances resolved at the level where the necessary variety exists. Failures arise when responsibility falls between levels, several levels intervene in the same problem, or lower-level autonomy is invoked to evade higher-level accountability.

4.7 Channels, variety, and algedonic escalation

Ashby's law of requisite variety states, in essence, that effective regulation requires a repertoire of responses adequate to the variety of disturbances that matter (Ashby 1958). Organizations cope by amplifying useful response variety and attenuating environmental variety through categorization, filtering, aggregation, and delegation. Communication channels are therefore active organizational mechanisms. They decide which differences survive long enough to influence action.

Agentic systems make this visible. Context windows, summary routines, access controls, message formats, routing rules, and human attention all constrain channel capacity. Compression is unavoidable; the issue is whether it removes noise or erases weak signals and minority interpretations. Relevant observations include latency, loss, distortion, compression ratios, unreported anomalies, and the relation between channel load and decision quality.

Algedonic channels provide exceptional escalation outside routine reporting. The term is often associated with pain or urgent threat, but unusually favorable developments can also require rapid attention. In hybrid systems, an algedonic mechanism could be triggered by model uncertainty, anomalous environmental feedback, safety thresholds, or a human participant's challenge. Its performance should be evaluated through detection-to-escalation time, false-positive rates, response appropriateness, and whether routine hierarchy can suppress the signal. The twin failures are silent catastrophe and alarm saturation.

4.8 A non-equivalence principle

The mapping above is intentionally non-isomorphic. A VSM function may be performed by several people and machines, one component may contribute to several functions, and some functions may be embedded in protocols or institutions rather than agents. That a function is carried by people, representations, and artifacts together rather than by any single component is the central finding of distributed-cognition research (Hutchins 1995). This non-equivalence principle prevents the framework from becoming a branding exercise. The empirical question is whether the function is performed effectively, not whether a diagram contains the expected labels.

5. Hard problems revealed by the synthesis

5.1 Recursive boundaries and membership

The boundary of an agentic system is often drawn around whichever software components its developer controls. That boundary may be convenient for

engineering but misleading for organizational analysis. If a human approval step, external database, procurement constraint, or downstream institution shapes the outcome, excluding it can conceal causal dependencies and relocate accountability.

A VSM analysis begins by naming the system-in-focus and its transformation. Researchers should then identify which units produce that transformation, what environments each encounters, and at which level particular disturbances can be regulated. Boundary choices should be treated as contestable hypotheses. Critical systems heuristics turns this into a method: every systems design rests on boundary judgments about whose purposes, expertise, and interests count, and because those judgments are normative rather than technical they must be exposed to argument between the involved and the affected (Ulrich 1987; Midgley 2000). A useful test is counterfactual: if the excluded element changed, could the system still maintain its identity and operations? If not, the analytical boundary probably requires justification or revision.

5.2 Collective identity and legitimate purpose

Multi-agent systems are often given a common objective and thereby assumed to possess collective purpose. This conflates specification with legitimacy. Objectives can be inconsistent across prompts and policies, altered through interaction, or operationalized in ways that affected people would reject. The gap between a stated objective and the outcome its designers intended is a standing problem in AI safety research (Amodei et al. 2016), and it does not close as systems improve: in reinforcement-learning environments with deliberately flawed reward functions, more capable agents earn more proxy reward and less true reward, with sharp transitions in behaviour at capability thresholds (Pan, Bhatia, and Steinhardt 2022). The alignment literature draws the distinction this argument needs, between aligning a system to its instructions and aligning it to principles that those affected could reflectively endorse (Gabriel 2020). Moreover, a system can remain internally coherent while pursuing harmful ends. Experiments in autonomous pricing agents illustrate the distinction: collective coherence may sustain supra-competitive outcomes that are operationally effective yet socially unacceptable (Fish et al. 2024).

The key question is not whether agents can state a shared mission but how binding purposes are authorized, interpreted, contested, and changed. Hybrid systems should preserve a trace from operational goals to human and institutional mandates. They should also expose disagreement rather than fabricating unanimity. This is a sharper requirement for language-model collectives than for human committees. Assistants trained on human prefer-

ence data are measurably sycophantic, and matching a user's stated position predicts a response being preferred (Sharma et al. 2024). That evidence concerns assistant-to-user behaviour rather than agreement among agents, but if the same preference dynamics operate between agents, apparent consensus is weak evidence of independent judgment. Research should compare arrangements in which purpose is centrally specified, deliberatively authorized, contractually distributed, or allowed to emerge from interaction, assessing both performance and legitimacy. Participatory arrangements should not be assumed to settle the legitimacy question by themselves, since participation can be co-opted into a legitimacy claim without any redistribution of power (Birhane et al. 2022).

5.3 Channel capacity and epistemic loss

The proliferation of agents can expand observation and response variety while simultaneously overwhelming the channels through which the collective makes sense of them. Summarization, voting, ranking, and managerial aggregation reduce load but can delete causal context, uncertainty, and dissent. Group research has measured the effect directly: discussion disproportionately surfaces what members already hold in common, so groups converge on their pre-discussion consensus even when the information held privately by individuals points elsewhere (Stasser and Titus 1985). The important variable is not message volume alone but the preservation of decision-relevant differences.

Experiments can manipulate channel bandwidth, compression, topology, and access to shared state while tracking anomaly detection, convergence quality, and recovery. Shannon's information theory supplies measures of capacity, noise, and distortion but is deliberately indifferent to semantic relevance (Shannon 1948). Relevance must therefore be evaluated against the operational task, the system's declared purpose, and stakeholder-defined harms. The VSM predicts that both insufficient attenuation and excessive attenuation are pathological. This produces a non-monotonic expectation: more information and connectivity need not improve viability.

5.4 Conflict mediation without premature consensus

Agent collectives require mechanisms for resolving incompatible recommendations. Simple majority voting, debate, or superior-agent arbitration can generate an answer without addressing whether disagreement reflects missing information, different values, or jurisdictional conflict. Multi-agent debate does not reliably outperform strong aggregation baselines and is sensitive to protocol choices (Smit et al. 2024). Social choice theory shows that part of the difficulty is formal rather than motivational: no aggrega-

tion rule that is universal, anonymous, and systematic can guarantee a consistent collective position across interdependent propositions, so a collective can endorse each premise by majority and reject the conclusion those premises support (List and Pettit 2002, 2011). Consensus can therefore be a failure mode rather than a success metric, and dissent a resource, since disagreement is what forces privately held information into the discussion (Schulz-Hardt et al. 2006).

System 2 highlights mediation and oscillation control; System 5 highlights conflicts that implicate identity. Research should distinguish operational disagreements that can be harmonized, resource conflicts requiring regulation, and normative conflicts requiring authorized policy judgment. Measures should include the quality and diversity of reasons retained after resolution, not merely time to consensus.

5.5 Exception handling and algedonic escalation

AI agents are often deployed because routine cases can be processed at scale. Viability is tested by cases for which the routine does not fit. An exception may be statistical novelty, rule conflict, ethically salient harm, or a situation in which the authorized goal itself is called into question. Systems that treat every exception as another classification problem risk normalizing what should interrupt the process. Organizational safety research has named that mechanism: at NASA, anomalous signals were repeatedly reinterpreted as acceptable within existing engineering rules until deviance became officially routine (Vaughan 1996). Where interactive complexity and tight coupling combine, such failures are a structural property of the design rather than an operator error (Perrow 1984).

Research should examine who can declare an exception, what evidence survives escalation, whether agents can block or dilute the signal, and which authority can change the governing rule. High-reliability organizations answer the first of these by migrating the decision to whoever holds the most situational knowledge, irrespective of rank, and by treating small failures as information rather than noise (Weick and Sutcliffe 2015). Monitoring and pre-action intervention have already improved outcomes in several agent environments (Barbi, Yoran, and Geva 2025), providing an adjacent experimental precedent. Explicit escalation may increase cost and false alarms; its value lies in preserving the organization's ability to respond before routine failure becomes systemic.

5.6 Distributed accountability

In hybrid collectives, one actor may frame a problem, another retrieve evidence, another recommend action, a human approve it, and a tool execute

it. A post hoc request for "the responsible agent" misunderstands the organizational chain. Political ethics described this long before autonomous systems, as the problem of many hands: where many actors each contribute marginally to an outcome, both the hierarchical excuse and the collective excuse are available, and each dissolves responsibility that ought to remain assignable (Thompson 1980; van de Poel et al. 2012). Learning systems sharpen the problem, because behaviour no operator could have predicted weakens the control-and-foreseeability condition on which responsibility ascription conventionally rests (Matthias 2004). Accountability requires traceability of contributions, authority, and opportunities to intervene.

Audit logs alone are insufficient if they record events without the institutional meaning of roles and mandates. Meaningful human control requires both responsiveness to relevant human reasons and traceability to humans who can understand and answer for system behavior (Santoni de Sio and van den Hoven 2018). A viable accountability architecture must show which decisions were delegated, what constraints applied, when uncertainty was known, who could contest the decision, and which level had authority to revise policy. Accountability is thus a recursive property: local responsibility must be related to the identity and governance of the larger systems that authorized the work.

6. A falsifiable research programme

The framework becomes scientifically useful only when its constructs can fail. The propositions below translate the synthesis into comparative hypotheses and specify observations that could count against the VSM theory. They are not presented as consequences already established by it.

6.1 Proposition 1: capability and organization interact

P1: Increasing constituent-agent capability will not reliably improve collective performance when coordination and escalation capacity remain fixed.

A factorial experiment could vary constituent capability and collective architecture while holding tools, budgets, and tasks constant. Agent-level competence could be established with an interactive benchmark such as AgentBench (Liu et al. 2024). Architectures could include a single agent, a flat peer network, a fixed orchestrator, and an adaptive coordinator. Outcomes should include task success, milestone completion, recovery from planted errors, constraint violations, communication cost, duplicated work, and between-run variance. The critical test is the interaction term: if stronger agents improve every architecture equally, the proposition is weakened;

if gains depend on the organization's capacity to coordinate and escalate, capability and collective viability are empirically separable.

6.2 Proposition 2: escalation supports recovery

P2: Collectives with explicit escalation mechanisms will recover more effectively from unmodeled disturbances than collectives relying only on peer coordination.

Experimental treatments could compare no escalation, fixed-threshold escalation, and escalation requiring human confirmation. After stable operation, researchers would inject disturbances absent from the agents' normal task specification. Measures would include detection precision and recall, time to escalation, correct routing, false-alarm burden, cumulative performance loss, time to recovery, and recurrence. The proposition would be challenged if escalation merely centralizes routine work, produces intolerable alarm saturation, or fails to improve recovery. Alarm saturation is not a hypothetical failure mode: clinical reviews report that between 72 and 99 per cent of hospital alarms are false, and link that volume to desensitization and to alarms being missed (Sendelbach and Funk 2013).

6.3 Proposition 3: attenuation has a viability frontier

P3: Central attenuation of local information will improve short-run coherence only up to a point, beyond which it reduces adaptive performance.

This proposition predicts a nonlinear relationship. Bandwidth, summarization ratio, routing topology, and the proportion of local observations visible centrally can be manipulated. Researchers should measure short-run agreement and throughput alongside local-anomaly recall, task-relevant information retained, response diversity, and adaptation after a regime shift. The information bottleneck offers one formal approach to evaluating compression that preserves task-relevant information (Tishby, Pereira, and Bialek 2000). A monotonic improvement from greater compression, or no reduction in novelty detection at extreme attenuation, would count against the proposed trade-off.

6.4 Proposition 4: present regulation and future intelligence should be differentiated

P4: Explicitly differentiating operational regulation from environmental intelligence will reduce goal drift under changing conditions.

A fused controller could be compared with an architecture that separates operational evaluation from environmental scanning and scenario generation, with an authorized policy process adjudicating conflicts. Outcomes would

include current-task performance, forecast calibration, change-detection delay, adaptation latency, exploration coverage, and deviation from an authorized charter. Organization theory motivates the distinction under the heading of ambidexterity. Tushman and O'Reilly (1996) argue that exploiting current operations and exploring future ones require structurally separate units held together by a common senior team; the subsequent review literature treats structural separation as only one realization, alongside contextual and sequential forms in which the same unit or the same individuals alternate between the two (Raisch and Birkinshaw 2008). Which realization suits a hybrid collective is therefore open rather than settled, and the VSM System 3–System 4 relationship has not been directly tested in agent collectives.

6.5 Proposition 5: authorization and traceability affect accountability

P5: Systems with contestable human authorization of identity constraints and complete decision provenance will produce more attributable and procedurally legitimate outcomes than systems in which purpose emerges only from agent interaction.

Experiments should separate the two parts of the claim. Technical attribution can be measured through trace completeness, independent reconstruction of decision chains, the identification of observation, recommendation, decision, and execution roles, and the success of overrides. Procedural legitimacy requires human or institutional evaluation: affected participants' understanding, perceived voice, contestability, and acceptance before and after outcomes are known. These are established constructs. Acceptance of a decision depends substantially on the perceived fairness of the procedure and on having had an opportunity to be heard, over and above how favourable the outcome was (Thibaut and Walker 1975; Folger 1977; Lind and Tyler 1988). Contestability has more recently been treated as an architectural property to be designed into an automated decision system rather than added to it afterwards (Almada 2019; Alfrink et al. 2023). A developer-authored prompt should not be treated as equivalent to authorization by a manager, professional body, affected community, or democratic institution. The relevant authorizing public must be specified for each demonstrator.

6.6 Proposition 6: recursion is conditionally beneficial

P6: Recursive governance arrangements will outperform flat architectures under multiscale disturbances only when escalation thresholds and jurisdictional boundaries are sufficiently clear.

A 2×2×2 design could vary flat versus recursive architecture, single-scale versus multiscale disturbance, and clear versus overlapping jurisdiction. Outcomes would include local and global task success, recovery time by

scale, escalation-routing accuracy, contradictory interventions, unowned problems, duplicated control, communication cost, and robustness to the loss of a coordinating node. The last of these matters most where the collective's communication topology is hub-dominated, since such networks tolerate the random loss of nodes well but fragment when their most connected nodes are removed (Albert, Jeong, and Barabási 2000). Institutional research supplies a precedent for the recursive arm of the comparison: among long-enduring common-pool-resource institutions, those that form part of larger systems organize appropriation, monitoring, enforcement, conflict resolution and governance in multiple nested layers rather than a single tier (Ostrom 1990). A simple hierarchy should not be counted as recursive unless its operational units themselves perform the relevant viability functions. The proposition would be weakened if recursion adds overhead without improving multiscale recovery, or if jurisdictional clarity has no moderating effect.

6.7 Three demonstrator domains

Emergency response. A simulation can compare flat, hierarchical, and recursive teams responding to localized and system-wide disruptions. This domain makes disturbance, escalation, time pressure, and cross-level effects visible, and it has a real organizational referent: field study of fire-department incident command shows how a formalized command structure supports reliable performance in volatile conditions by reconfiguring roles and authority as an incident develops (Bigley and Roberts 2001). Its weakness is that simulated authority and identity remain researcher-defined.

Hybrid public service. A bounded administrative workflow can examine how AI recommendations, human authorization, audit, contestability, and exceptional cases interact. The exceptional case is the point of interest. Unlike a street-level bureaucrat, who can refine the decision criteria on encountering a case the rules do not fit, an algorithmic decision system must decide first and revise only afterwards, so the cost of the mismatch falls on the person in front of it (Alkhatib and Bernstein 2019). This domain offers high external validity for accountability, but only low-risk or sandboxed decisions should be used until the architecture is understood.

Distributed knowledge work. Scientific or software-production teams can be organized with different coordination, verification, and strategic-learning functions. Formally credited knowledge production has already shifted decisively from individuals to teams (Wuchty, Jones, and Uzzi 2007), and distributed teams are measurably slower than co-located ones, with the delay traceable to the number of people a change involves and the difficulty of reaching the right person across sites (Herbsleb and Mockus 2003).

This domain permits rich trace data and repeated tasks, although project completion is still only a proxy for long-term organizational viability.

Across all three, final accuracy should be supplemented by trajectory measures. Evaluation of agentic systems has been criticized on exactly this ground, for attending to final outcomes while ignoring the step-by-step character of the process that produced them (Zhuge et al. 2025). Table 1 summarizes the translation from VSM concepts to empirical observations.

Table 1. Functional constructs, observations, and characteristic failures

Construct	Agent-collective problem	Candidate observations	Characteristic failure
System 1	Preserve useful local autonomy while producing a collective outcome	local task success, autonomy retained, inter-unit dependencies	brittle centralization; incoherent local optimization
System 2	Mediate dependencies and damp oscillation	conflict duration, duplicate work, message churn, convergence diversity	deadlock; thrashing; premature consensus
System 3/3*	Allocate resources, enforce constraints, and verify state	audit coverage, reallocation, discrepancies, verification accuracy	fabricated status; uncorrected drift; micromanagement
System 4	Model environmental change and generate alternatives	adaptation latency, scenario diversity, novelty detection, exploration	reactive optimization; stale models; lock-in
System 5	Maintain authorized identity and resolve present–future trade-offs		

charter consistency, authorization trace, policy overrides, contestability

goal drift; illegitimate purpose; paralysis

Recursion

Match autonomy and governance to multiple scales

routing accuracy, jurisdiction clarity, cross-level contradiction

gaps, duplication, or contradictory control

Channels/variety

Preserve decision-relevant differences under limited capacity

latency, loss, distortion, relevant information retained

overload; over-attenuation; epistemic blindness

Algedonic escalation

Route exceptional threats and opportunities

escalation delay, precision/recall, recovery

silent catastrophe; alarm saturation

7. Discussion: contribution, alternatives, and limits

The proposed framework belongs among several established approaches rather than above them. Multi-agent systems engineering provides formal and experimental methods for coordination, communication, task allocation, and distributed decision-making. Agent-based modelling offers generative explanations of macro-patterns from specified micro-interactions (Epstein 1999; Bonabeau 2002). Collective-intelligence research demonstrates that group performance can be a property not reducible to the intelligence of individual members (Woolley et al. 2010). Distributed cognition locates cognitive accomplishment in relations among people, representations, and artifacts (Hutchins 1995). Sociotechnical systems theory insists that technical and social organization must be designed jointly (Trist and Bamforth 1951). Institutional analysis foregrounds rules, boundaries, monitoring, collective choice, and polycentric governance (Crawford and Ostrom 1995; Ostrom 2010).

The VSM's distinctive value is integrative. It connects operational autonomy, anti-oscillatory coordination, internal regulation and audit, future-oriented intelligence, identity and policy, recursion, and exceptional escalation in one account of continued organization. This produces questions that an interaction protocol alone does not answer: Which collective is being made

viable? At what level? Around whose purposes? Which differences must its channels preserve? Who may interrupt routine operation? How can the organization revise its models without allowing unaccountable goal change?

That integration also creates risks. First, the VSM can be made unfalsifiable if every successful function is retrospectively relabeled as one of its systems. The non-equivalence principle and advance specification of measures are intended to prevent this. Second, the VSM's claims of necessary and sufficient functions are under-tested empirically. One quantitative study of start-ups produced mixed component-level results rather than decisive confirmation (Crisan Tran 2008). The present article therefore treats the model as a disciplined source of hypotheses, not a validated universal architecture.

Third, most agent benchmarks are not persistent organizations. They have explicit rewards, compressed timescales, researcher-defined boundaries, and limited consequences. Real organizations contain contested purposes, legal obligations, power asymmetries, professional norms, affected nonparticipants, and histories that cannot be reset between trials. Demonstrator results must not be generalized from benchmark coordination to institutional viability without field evidence.

Fourth, viability is not legitimacy. Autonomous agents can coordinate effectively toward collusion or harm. Human authorization is also not a magic phrase: approval by a developer, executive, frontline professional, affected population, or democratic institution carries different authority. The paper's normative position is therefore procedural. Consequential collective purposes should remain connected to identifiable human institutions with legitimate authority, meaningful opportunities for contestation, and traceable responsibility. Which institutions qualify is a domain-specific political question, not something the VSM can decide by itself.

Finally, the organizational frame should not erase material constraints. Model ownership, computation, labor, data access, environmental cost, and platform power determine which varieties can be sensed and which actions are available. A formally complete VSM mapping could still conceal exploitative or brittle infrastructures. Sociotechnical and institutional analysis remain necessary correctives.

Within these limits, the framework makes a useful shift in research design. It replaces the binary comparison "one agent or many?" with structured comparisons of collective functions. It directs evaluation toward recovery, adaptation, information loss, jurisdiction, identity, and answerability. It also turns the VSM–agent connection into a programme that evidence can reject, revise, or bound.

8. Conclusion

The central challenge of agentic AI is no longer only how to construct a capable autonomous agent, but how to organize populations of human and synthetic actors so that they can act coherently without destroying useful autonomy, adapt without drifting from legitimate purpose, and distribute work without dissolving responsibility.

The Viable System Model offers a functional vocabulary for this challenge. Used cautiously, it directs attention to the five functions of operation, coordination, regulation, intelligence and identity, and to the recursion, channels and exceptional escalation that connect them. It does not prove that every agent collective is a viable system or prescribe a universal architecture. Its scientific value lies in the questions it makes precise and the failures it makes observable.

The next step is empirical comparison. If VSM-derived architectures do not improve recovery, adaptation, information preservation, or accountability under matched conditions, the proposed synthesis should be narrowed or rejected. If they do, organizational cybernetics supplies something agent research currently lacks: an account of how increasingly capable agents can be organized into collectives that stay governable, adaptive, and answerable.

References

Albert, R., Jeong, H., & Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature, 406*(6794), 378–382. https://doi.org/10.1038/35019019

Alfrink, K., Keller, I., Kortuem, G., & Doorn, N. (2023). Contestable AI by design: Towards a framework. *Minds and Machines, 33*(4), 613–639. https://doi.org/10.1007/s11023-022-09611-z

Alkhatib, A., & Bernstein, M. (2019). Street-level algorithms: A theory at the gaps between policy and decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Paper 530, pp. 1–13). ACM. https://doi.org/10.1145/3290605.3300760

Almada, M. (2019). Human intervention in automated decision-making: Toward the construction of contestable systems. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law* (pp. 2–11). ACM. https://doi.org/10.1145/3322640.3326699

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv*. https://arxiv.org/abs/1606.06565

Ardavs, A., Pudane, M., Lavendelis, E., & Nikitenko, A. (2019). Long-term adaptivity in distributed intelligent systems: Study of ViaBots in a simulated environment. *Robotics, 8*(2), 25. https://doi.org/10.3390/robotics8020025

Ashby, W. R. (1958). Requisite variety and its implications for the control of complex systems. *Cybernetica, 1*(2), 83–99. Primary text

Bakan, D. (1966). *The duality of human existence: An essay on psychology and religion*. Rand McNally.

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual Review of Psychology, 52*(1), 1–26. https://doi.org/10.1146/annurev.psych.52.1.1

Barbi, O., Yoran, O., & Geva, M. (2025). Preventing rogue agents improves multi-agent collaboration. In *Proceedings of the REALM Workshop* (pp. 486–511). https://aclanthology.org/2025.realm-1.34/

Bavelas, A. (1950). Communication patterns in task-oriented groups. *The Journal of the Acoustical Society of America, 22*(6), 725–730. https://doi.org/10.1121/1.1906679

Beer, S. (1979). *The heart of enterprise*. Wiley.

Beer, S. (1981). *Brain of the firm* (2nd ed.). Wiley.

Beer, S. (1984). The Viable System Model: Its provenance, development, methodology and pathology. *Journal of the Operational Research Society, 35*(1), 7–25. https://doi.org/10.1057/jors.1984.2

Beer, S. (1985). *Diagnosing the system for organizations*. Wiley.

Beer, S. (2002). What is cybernetics? *Kybernetes, 31*(2), 209–219. https://doi.org/10.1108/03684920210417283

Bigley, G. A., & Roberts, K. H. (2001). The incident command system: High-reliability organizing for complex and volatile task environments. *Academy of Management Journal, 44*(6), 1281–1299. https://doi.org/10.2307/3069401

Birhane, A., Isaac, W., Prabhakaran, V., Diaz, M., Elish, M. C., Gabriel, I., & Mohamed, S. (2022). Power to the people? Opportunities and challenges for participatory AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1–8). ACM. https://doi.org/10.1145/3551624.3555290

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences, 99*(suppl. 3), 7280–7287. https://doi.org/10.1073/pnas.082080899

Castelfranchi, C. (1995). Guarantees for autonomy in cognitive agent architecture. In M. J. Wooldridge & N. R. Jennings (Eds.), *Intelligent agents:*

ECAI-94 workshop on agent theories, architectures, and languages (Lecture Notes in Computer Science, Vol. 890, pp. 56–70). Springer. https://doi.org/10.1007/3-540-58855-8_3

Cemri, M., Pan, M. Z., Yang, S., Agrawal, L. A., Chopra, B., Tiwari, R., et al. (2025). Why do multi-agent LLM systems fail? *arXiv*. https://arxiv.org/abs/2503.13657

Chan, A., Ezell, C., Kaufmann, M., Wei, K., Hammond, L., Bradley, H., et al. (2024). Visibility into AI agents. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 958–973). ACM. https://doi.org/10.1145/3630106.3658948

Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., et al. (2023). Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 651–666). ACM. https://doi.org/10.1145/3593013.3594033

Crawford, S. E. S., & Ostrom, E. (1995). A grammar of institutions. *American Political Science Review, 89*(3), 582–600. https://doi.org/10.2307/2082975

Crisan Tran, C. I. (2008). Assessing the Viable System Model: An empirical test of the viability-hypothesis. *International Journal of Applied Systemic Studies, 2*(1/2), 66–81. https://doi.org/10.1504/IJASS.2008.022795

Daft, R. L., & Weick, K. E. (1984). Toward a model of organizations as interpretation systems. *Academy of Management Review, 9*(2), 284–295. https://doi.org/10.2307/258441

Emirbayer, M., & Mische, A. (1998). What is agency? *American Journal of Sociology, 103*(4), 962–1023. https://doi.org/10.1086/231294

Epstein, J. M. (1999). Agent-based computational models and generative social science. *Complexity, 4*(5), 41–60. https://doi.org/10.1002/(SICI)1099-0526(199905/06)4:5%3C41::AID-CPLX9%3E3.0.CO;2-F

Espejo, R., & Reyes, A. (2011). *Organizational systems: Managing complexity with the Viable System Model*. Springer. https://doi.org/10.1007/978-3-642-19109-1

Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2024). Algorithmic collusion by large language models. *arXiv*. https://arxiv.org/abs/2404.00806

Folger, R. (1977). Distributive and procedural justice: Combined impact of "voice" and improvement on experienced inequity. *Journal of Personality and Social Psychology, 35*(2), 108–119. https://doi.org/10.1037/0022-3514.35.2.108

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines, 30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Herbsleb, J. D., & Mockus, A. (2003). An empirical study of speed and communication in globally distributed software development. *IEEE Transactions on Software Engineering, 29*(6), 481–494. https://doi.org/10.1109/TSE.2003.1205177

Herrera, C., Belmokhtar Berraf, S., & Thomas, A. (2012). Viable System Model approach for holonic product-driven manufacturing systems. In *Service Orientation in Holonic and Multi-Agent Manufacturing Control* (pp. 169–181). Springer. https://doi.org/10.1007/978-3-642-27449-7_13

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., & Zhou, D. (2024). Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=IkmD3fKBPQ

Hutchins, E. (1995). *Cognition in the wild*. MIT Press. Publisher record

Kuhr, L., Mennenga, M., & Herrmann, C. (2026). Cognitive AI agents in life cycle management of Industry 5.0 organizations: A conceptual framework. *Procedia CIRP, 140*, 1175–1180. https://doi.org/10.1016/j.procir.2026.05.198

Lawrence, P. R., & Lorsch, J. W. (1967). Differentiation and integration in complex organizations. *Administrative Science Quarterly, 12*(1), 1–47. https://doi.org/10.2307/2391211

Lind, E. A., & Tyler, T. R. (1988). *The social psychology of procedural justice*. Plenum Press. https://doi.org/10.1007/978-1-4899-2115-4

List, C., & Pettit, P. (2002). Aggregating sets of judgments: An impossibility result. *Economics and Philosophy, 18*(1), 89–110. https://doi.org/10.1017/S0266267102001098

List, C., & Pettit, P. (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199591565.001.0001

Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., et al. (2024). AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations*. Official proceedings

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences, 108*(22), 9020–9025. https://doi.org/10.1073/pnas.1008636108

Macal, C. M., & North, M. J. (2010). Tutorial on agent-based modelling and simulation. *Journal of Simulation, 4*(3), 151–162. https://doi.org/10.1057/

jos.2010.3

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science, 2*(1), 71–87. https://doi.org/10.1287/orsc.2.1.71

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

Midgley, G. (2000). *Systemic intervention: Philosophy, methodology, and practice*. Kluwer Academic/Plenum. https://doi.org/10.1007/978-1-4615-4201-8

Milgram, S. (1974). *Obedience to authority: An experimental view*. Harper & Row.

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press. https://doi.org/10.1017/CBO9780511807763

Ostrom, E. (2010). Beyond markets and states: Polycentric governance of complex economic systems. *American Economic Review, 100*(3), 641–672. https://doi.org/10.1257/aer.100.3.641

Pan, A., Bhatia, K., & Steinhardt, J. (2022). The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=JYtwGwIL7ye

Pérez Ríos, J. (2025). The Viable System Model and the taxonomy of organizational pathologies in the age of artificial intelligence (AI). *Systems, 13*(9), 749. https://doi.org/10.3390/systems13090749

Perrow, C. (1984). *Normal accidents: Living with high-risk technologies*. Basic Books.

Power, M. (1997). *The audit society: Rituals of verification*. Oxford University Press.

Qin, Y., Lee, R. T., & Sajda, P. (2026). Covert perception of AI adversely impacts team performance and changes physiological dynamics despite human-level AI competence. *npj Artificial Intelligence*. https://doi.org/10.1038/s44387-026-00130-1

Raisch, S., & Birkinshaw, J. (2008). Organizational ambidexterity: Antecedents, outcomes, and moderators. *Journal of Management, 34*(3), 375–409. https://doi.org/10.1177/0149206308316058

Rammert, W. (2008). *Where the action is: Distributed agency between humans, machines, and programs* (TUTS Working Papers 4-2008). Technische Univer-

sität Berlin. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-12331

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

Schulz-Hardt, S., Brodbeck, F. C., Mojzisch, A., Kerschreiter, R., & Frey, D. (2006). Group decision making in hidden profile situations: Dissent as a facilitator for decision quality. *Journal of Personality and Social Psychology, 91*(6), 1080–1093. https://doi.org/10.1037/0022-3514.91.6.1080

Sendelbach, S., & Funk, M. (2013). Alarm fatigue: A patient safety concern. *AACN Advanced Critical Care, 24*(4), 378–386. https://doi.org/10.4037/NCI.0b013e3182a903f9

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379–423, 623–656. Part I; Part II

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., et al. (2024). Towards understanding sycophancy in language models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=tvhaxkMKAn

Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., et al. (2023). *Practices for governing agentic AI systems* [White paper]. OpenAI. https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf

Smit, A. P., et al. (2024). Should we be going MAD? A look at multi-agent debate strategies for LLMs. In *Proceedings of the 41st International Conference on Machine Learning* (pp. 45883–45905). https://proceedings.mlr.press/v235/smit24a.html

Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology, 48*(6), 1467–1478. https://doi.org/10.1037/0022-3514.48.6.1467

Suchman, M. C. (1995). Managing legitimacy: Strategic and institutional approaches. *Academy of Management Review, 20*(3), 571–610. https://doi.org/10.2307/258788

Thibaut, J., & Walker, L. (1975). *Procedural justice: A psychological analysis*. Lawrence Erlbaum Associates.

Thompson, D. F. (1980). Moral responsibility of public officials: The problem of many hands. *American Political Science Review, 74*(4), 905–916. https://doi.org/10.2307/1954312

Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. *arXiv:physics/0004057*. https://arxiv.org/abs/physics/0004057

Trist, E. L., & Bamforth, K. W. (1951). Some social and psychological consequences of the Longwall method of coal-getting. *Human Relations, 4*(1), 3–38. https://doi.org/10.1177/001872675100400101

Tushman, M. L., & O'Reilly, C. A. (1996). Ambidextrous organizations: Managing evolutionary and revolutionary change. *California Management Review, 38*(4), 8–30. https://doi.org/10.2307/41165852

Ulrich, W. (1987). Critical heuristics of social systems design. *European Journal of Operational Research, 31*(3), 276–283. https://doi.org/10.1016/0377-2217(87)90036-1

van de Poel, I., Nihlén Fahlquist, J., Doorn, N., Zwart, S., & Royakkers, L. (2012). The problem of many hands: Climate change as an example. *Science and Engineering Ethics, 18*(1), 49–67. https://doi.org/10.1007/s11948-011-9276-0

Vaughan, D. (1996). *The Challenger launch decision: Risky technology, culture, and deviance at NASA*. University of Chicago Press.

Weick, K. E., & Sutcliffe, K. M. (2015). *Managing the unexpected: Sustained performance in a complex world* (3rd ed.). Wiley.

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

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science, 330*(6004), 686–688. https://doi.org/10.1126/science.1193147

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science, 316*(5827), 1036–1039. https://doi.org/10.1126/science.1136099

Zhuge, M., Zhao, C., Ashley, D. R., Wang, W., Khizbullin, D., Xiong, Y., et al. (2025). Agent-as-a-judge: Evaluate agents with agents. In *Proceedings of the 42nd International Conference on Machine Learning* (PMLR Vol. 267, pp. 80569–80611). https://proceedings.mlr.press/v267/zhuge25a.html

Zomer, N., & De Domenico, M. (2026). Unraveling the emergence of collective behavior in networks of cognitive agents. *npj Artificial Intelligence, 2*, 36.

https://doi.org/10.1038/s44387-026-00091-5

