

From Tools to Actors: How Agentic AI Reshapes Organizational Learning Dynamics

Completed Research Paper

Amirsiavosh Bashardoust

University of Lausanne - Faculty of
Business and Economics (HEC
Lausanne)
Quartier de Chamberonne, CH-1015,
Lausanne, Switzerland
Amirsiavosh.bashardoust@unil.ch

Kazem Haki

University of St.Gallen
Dufourstrasse 50, 9000 St. Gallen,
Switzerland
kazem.haki@unisg.ch

Yash Raj Shrestha

University of Lausanne - Faculty of
Business and Economics (HEC
Lausanne)
Quartier de Chamberonne, CH-1015,
Lausanne, Switzerland
Yashraj.shrestha@unil.ch

Abstract

Agentic AI evolves AI from a tool that supports human work into a delegated organizational actor capable of goal-directed reasoning, adaptive self-regulation, and consequential action-taking. This shift raises a central question: how does agentic AI reshape the dynamics through which organizations accumulate knowledge while preserving belief diversity? Building on and extending March's (1991) mutual-learning model, we develop an agent-based model that introduces agentic AI as a new class of learner-actor. We examine how its key features affect knowledge accumulation and belief diversity under varying levels of environmental turbulence. Our findings show that agentic AI's consequential action-taking acts as a turbulence-contingent lever, driving both knowledge and diversity only under unstable conditions. Human reconfiguration of agentic AI sustains diversity in stable environments but becomes redundant under turbulence. In contrast, agentic AI's self-regulation introduces a knowledge-diversity trade-off that weakens as turbulence increases. These findings contribute to IS and organizational learning research.

Keywords: Agentic AI, organizational learning, knowledge accumulation, belief diversity, agent-based modeling

Introduction

Agentic Artificial Intelligence (Agentic AI) systems are increasingly being embedded in organizational workflows and routines as delegated actors capable of perceiving tasks, formulating plans, using tools, and executing actions with limited human oversight. Industry reports suggest that organizations are rapidly experimenting with such systems, with 97% of executives reporting widespread deployment of AI agents in the past year and that roughly half of employees interact with them regularly (Writer, 2026). In parallel, Gartner forecasts that 40% of enterprise applications will embed task-specific agents by 2026, up from less

than 5% in 2025. This development marks an important shift in the organizational role of AI: from a technology that supports human decision-making to one that can participate directly in organizational action. The emerging literature refers to this class of systems as agentic AI: AI systems that can pursue goals, decompose tasks, adapt strategies, and act in organizational environments with varying degrees of autonomy (Humberd & Latham, 2026; Sapkota et al., 2025; Schneider, 2025; Shavit et al., 2023).

Theoretically, this shift matters because it challenges a foundational assumption in organizational learning: that organizational knowledge changes through human-mediated codification of individual beliefs into rules, routines, and procedures. Prior generations of AI such as rule-based systems, predictive machine learning (ML), and generative models, were integrated into organizations as instruments of human action: humans formulated goals, supplied data, interpreted outputs, and decided which beliefs earned a place in the organization's rules, routines, and procedures (Baird & Maruping, 2021; Berente et al., 2021; Faraj et al., 2018; Haki et al., 2025; Lebovitz et al., 2021). Agentic AI disturbs this established division of labor. Because it can formulate sub-goals, iteratively adapt its learning strategies through self-evaluation, and enact changes to organizational routines through direct execution rather than human mediation, it operates less as a passive tool and more as a delegated organizational actor with the capability that can change its environment beyond the normal routines in the organization (Baird & Maruping, 2021; Humberd & Latham, 2026).

These changes directly affect the core mechanisms of organizational learning. In March's (1991) seminal work, *organizational knowledge* emerges from mutual learning between members and an organizational code, which embodies and is continuously updated by the organization's shared beliefs about its environment. The long-run effectiveness of this process depends not only on *knowledge accumulation*, but also on the preservation of *belief diversity*. Organizations must preserve heterogeneity of beliefs long enough to explore the space of possible realities before converging on a shared view, as premature convergence lock them into inferior routines (Denrell & March, 2001; Levinthal & March, 1993; March, 1991). Subsequent information systems (IS) research has extended and refined this perspective by conceptualizing ML as an additional class of learner alongside humans, thereby reshaping the returns to human exploration and altering learning dynamics under conditions of environmental turbulence (Sturm et al., 2021).

However, existing organizational-learning models (e.g., Sturm et al., 2021) continue to treat AI as a passive learner or decision-support input whose effects are ultimately mediated by humans. In these models, AI learns from data, is configured by humans, and contributes to the organization only through channels that humans mediate—whether as a tool in the hands of a human learner or as a passive learner whose outputs are codified by humans into the organizational code. Agentic AI disrupts this assumption along three critical mechanisms. First, it can endogenously regulate and adapt its own learning, making its capabilities a function of internal evaluation rather than external human reconfigurations (Shinn et al., 2023; Sturm et al., 2021). Second, it can directly alter organizational routines through consequential action, thereby bypassing the traditional codification processes through which beliefs enter the organizational code (Baird & Maruping, 2021; Humberd & Latham, 2026; March, 1991). Third, because agents can self-regulate, the marginal returns to human oversight and intervention are fundamentally altered, such that human effort is no longer the sole adaptive force shaping AI behavior (Liu et al., 2025; Raisch & Krakowski, 2021; Sturm et al., 2021).

These mechanisms create theoretically ambiguous implications for organizational learning. Agentic AI may expand belief diversity by injecting independently generated perspectives at scale, or compress it by rapidly imposing self-generated beliefs on the organizational code. It may accelerate adaptation to shifting realities, or it may entrench drift in directions humans cannot readily detect. Thus, agentic AI may simultaneously enhance short-run learning and undermine the heterogeneity on which long-run adaptive capacity depends. Which of these effects prevails, and under what conditions, remains an open question. We thus ask two research questions:

RQ1. *How do the constitutive features of agentic AI reshape the tradeoff between organizational knowledge accumulation and belief diversity?*

RQ2. *How do these effects vary under environmental turbulence, when the underlying reality to which organizational beliefs must adapt is itself shifting?*

To address these questions, we develop an agent-based model (ABM) that extends March's (1991) mutual-learning framework by introducing agentic AI as a new class of learner-actor. Our model represents agentic AI not merely as an additional source of beliefs, but as an actor capable of generating, revising, and enacting beliefs through mechanisms that differ from both human learners and passive ML systems. Through simulation experiments, we examine how agentic AI affects *organizational knowledge level (accumulation)* and *belief diversity* under varying levels of environmental turbulence.

This study makes three contributions. First, we extend organizational learning theory by theorizing agentic AI as a new class of organizational learner-actor whose beliefs can be generated, revised, and enacted without direct human mediation. Second, we contribute to IS research on AI in organizations by specifying the mechanisms through which agentic AI differs from tool-like, generative, and passive ML systems. Third, we contribute to research on human-AI organizing by showing how agentic AI changes the value of human exploration and belief diversity under environmental turbulence. In doing so, the study offers guidance for organizations seeking to benefit from agentic AI while preserving the diversity of knowledge on which long-term learning depends.

Research Background

This study situates at the intersection of two literatures whose core assumptions are currently being revisited. IS and organization-theoretic work on AI has progressed from treating AI as a tool to conceptualizing it as a learner (Raisch & Krakowski, 2021), and more recently to recognizing it as a participant in organizational action (Humberd & Latham, 2026; Vanneste & Puranam, 2024). Organizational-learning theory, anchored in March's (1991) mutual-learning model, has progressively added technology into its formal machinery, most recently as a passive machine learner coupled to human learners (Sturm et al., 2021) or as a statistical selector of routine variants (Balasubramanian et al., 2022). Agentic AI challenges a foundational assumption shared by these streams: that the organizational code is a passive aggregator of beliefs contributed by human agents and processed by technology. An agent that sets its own sub-goals, acts directly on the environment, and adapts its behavior through self-evaluation (Humberd & Latham, 2026; Sapkota et al., 2025; Schneider, 2025) transforms the organizational code into a site of generation rather than aggregation. In what follows, we develop this claim in three steps.

March's (1991) Mutual-Learning Model and Its Simulation Lineage

March (1991) modelled organizations as mutual-learning systems in which individuals and an organizational code adapt to one another. The organizational code is the codified layer of shared beliefs—the procedures, routines, and accumulated lessons that persist beyond any single member and guide collective action. It is continuously updated through reciprocal learning with individuals as the environment changes. March (1991) showed that long-run performance depends on managing the trade-off between exploration (search for new alternatives, which maintains belief diversity) and exploitation (refinement of current alternatives, which homogenizes beliefs). His central premise is that adaptive processes refine exploitation more rapidly than exploration and can therefore become self-destructive. This finding anchors subsequent work on ambidexterity, myopia, and competence traps (Denrell & March, 2001; Lavie et al., 2010; Levinthal & March, 1993). In this tradition, belief diversity is not a residual of imperfect learning but a strategic asset that enables organizations to avoid premature lock-in.

A distinct simulation lineage has extended this framework. Miller et al. (2006) added interpersonal learning and tacit knowledge, establishing the pattern of introducing new learner classes while preserving March's formal core. Fang et al. (2010) introduced structural isolation of subgroups, showing that delayed belief convergence raises long-run knowledge accumulation (level). Posen and Levinthal (2012) demonstrated, against intuition, that environmental turbulence does not monotonically favor exploration: in moving environments, accumulated knowledge itself becomes obsolete, so the value of further search depends jointly on turbulence magnitude and the pace of exploitation. Kane and Alavi (2007) brought the model into IS by showing that different IT-enabled learning mechanisms favor exploration or exploitation differently and that environmental turbulence alters which mechanism pays off. Csaszar and Levinthal (2016) as well as Puranam and Swamy (2016) deepened the representational layer, showing how cognitive heterogeneity between coupled learners shapes the discovery of new strategies. Together these studies define a canonical way to extend March: introduce a new learner class or learning channel, specify how it

interacts with the organizational code, and trace the effect on knowledge accumulation (i.e., knowledge level) and belief diversity.

Within this lineage, AI has been introduced but only as a passive learner. Sturm et al. (2021) extend March's model by adding ML agents and show that high-capability ML can reduce the need for human exploration, human reconfiguration of ML exhibits nonlinear returns, and rapid codification of ML outputs becomes more valuable under environmental turbulence. Balasubramanian et al. (2022) argue that ML-driven selection based on historical data can reduce routine diversity—the behavioral counterpart to belief diversity—and exacerbate learning myopia. Hendriks et al. (2024) distinguish between AI learning and AI enactment, while Choudhary et al. (2025) formalize hybrid human–AI search processes. Despite these advances, a critical limitation remains: existing models do not represent AI as an autonomous agent. In prior work, AI systems learn from data but do not initiate action, set sub-goals, or coordinate behavior independently. Their role is confined to contributing inputs into human-mediated learning processes. In Floridi's (2025) terms, these systems are artifactual rather than agentic forms of artificial agency.

AI as Tool, Learner, and Participant

Scholars have long treated AI as a tool that supports human work (Alavi & Leidner, 2001; Iyengar et al., 2015). ML extended this view by enabling systems to derive patterns from data and exceed human performance on narrow tasks (Jordan & Mitchell, 2015). However, ML remains fundamentally operator-dependent: models are trained for specific problems, make predictions within those scopes, and require humans to set objectives, curate data, and redeploy them (Sturm et al., 2021). Generative AI extended the range of outputs to include novel text, code, or images, but continues to operate reactively in response to human prompts (Haki et al., 2025; Schneider, 2025).

Agentic AI marks a qualitatively different shift. Building on Large Language Models (LLMs) as reasoning engines, agentic systems combine perception, planning, tool use, and memory to pursue multi-step goals with limited human oversight (Sapkota et al., 2025; Wang et al., 2024; Xi et al., 2025). Recent taxonomies characterize them along dimensions such as autonomy, efficacy, goal complexity, and generality (Kasirzadeh & Gabriel, 2025). Although we have evidence from the literature which they considered AI as an active agent (Strich et al., 2021), however we define the distinction that matters is therefore not active versus passive AI, but which kind of action is delegated: task execution within human-designed routines, or the generation and enactment of routines themselves. Although Strich et al. (2021), assume traditional AI as an active agent in a sense that it could execute human-designed decision routines, Agentic AI would push the boundaries of autonomy further and enables itself to change its environment beyond the existent routines and processes. Floridi (2025) conceptualizes the resulting entity as artificial agency, categorically distinct from both human agents and prior artifactual systems.

The literature on artificial agency has converged on similar distinctions through diverse theoretical lenses. Baird and Maruping (2021) theorize agentic artifacts as bidirectional delegation partners, reframing IS use around who delegates what to whom. Murray et al. (2021) map four forms of conjoined human–technology agency and identify automating configurations in which technology exercises intentionality over both protocol development and action selection. Raisch and Krakowski (2021) articulate this as an automation–augmentation paradox whose resolution is temporal and scale-dependent. Vanneste and Puranam (2024) argue that perceived agency, rather than any objective autonomy threshold, is what shapes trust and delegation decisions. At the field level, Berente et al. (2021) frame autonomy, learning, and inscrutability as the three properties that make AI management distinctive, while Lyytinen et al. (2021) describe the resulting configurations as metahuman systems. Humberd and Latham (2026) extend these arguments through an agency-theoretic lens, and Stelmaszak et al. (2026) complement them with a relational view in which AI's effects emerge from human–algorithm codependence. What connects these accounts is a shared conclusion: sufficiently autonomous AI is not a faster ML system but a new participant in organizational action. What remains underexplored, however, is how organizational learning dynamics evolve in the presence of such participants.

Agentic AI Features

Agentic AI is best understood not as a new technology but as a recombination of capabilities that, taken together, cross a qualitative threshold, whereby AI artifact transitions from an instrument wielded by

human agents to a delegated actor within a sociotechnical system (Baird & Maruping, 2021; Humberd & Latham, 2026). Drawing on prior conceptualizations across disciplines, especially philosophy (Floridi, 2025) and IS (Berente et al., 2021; Schneider, 2025), we derive three features that jointly distinguish agentic AI from generative and traditional AI systems.

Feature 1: Autonomous goal-directed reasoning. The first distinguishing feature is the capacity to pursue goals through deliberation rather than through predefined rules or narrow statistical mappings, and to do so across multiple related dimensions of a problem domain (Sapkota et al., 2025; Xi et al., 2025). Whereas classical ML systems optimize a fixed objective over a single task dimension (Sturm et al., 2021), agentic AI reasons about which sub-goals to pursue, in what order, and under what environmental conditions (Wang et al., 2024). Schneider (2025) identifies this shift as the key discontinuity between generative AI—which produces outputs conditioned on prompts—and agentic AI, which reasons about action sequences with reference to environmental feedback. Two consequences follow for how agentic AI departs from ML. First, the scope of reasoning is no longer constrained to a single task dimension; a single agent may reason simultaneously about related aspects of a problem domain (Sapkota et al., 2025). Second, the agent can decompose a high-level goal into sub-goals and revise the decomposition as execution unfolds, without requiring a human to pre-structure the problem (Shinn et al., 2023; Wei et al., 2022). This is a capability that is impossible for narrow-task ML agents and underdeveloped in generative AI systems that operate in a single forward pass.

Feature 2: Adaptive self-regulation. The second feature is the agent’s capacity to monitor its own performance, evaluate outcomes against goals, and adjust its reasoning or learning strategy accordingly (Sapkota et al., 2025). Agents that evaluate their own outputs and retain those evaluations across episodes outperform otherwise identical systems that lack self-evaluation (Shinn et al., 2023). Crucially, self-regulation is not equivalent to learning in the classical ML sense. A learning algorithm updates parameters according to an externally specified loss; a self-regulating agent decides when to update, what to update, and how to evaluate whether an update succeeded (Floridi, 2025; Humberd & Latham, 2026; Sapkota et al., 2025).

Feature 3: Consequential action-taking. The third and most consequential feature is the capacity to take actions that materially alter the environment, including executing code, calling application programming interfaces, initiating transactions, and modifying organizational workflows (Baird & Maruping, 2021; Humberd & Latham, 2026). This is the feature that most sharply separates agentic AI from other AI paradigms. Generative AI produces artifacts but does not act on them; expert systems reason about actions but do not execute in live environments (Sapkota et al., 2025; Schneider, 2025). This feature’s implication is that agentic AI does not merely contribute beliefs that humans may or may not codify; it reshapes the environment in which other agents—human and artificial—subsequently form their beliefs (Baird & Maruping, 2021; Humberd & Latham, 2026). This feature bears most directly on the concerns raised by Balasubramanian et al. (2022), who argue that ML-based selection reduces routine diversity and deepens learning myopia because it privileges statistically dominant patterns from historical data. Consequential action-taking shares the diversity-threatening property — it lets machine-held beliefs enter the code without the deliberative filter of human codification — but differs in mechanism: whereas ML selection chooses among *existing* variants and is therefore inherently variance-reducing, an acting agent *enacts newly generated* beliefs, which may inject variation rather than suppress it.

Research Gap

March’s organizational code is updated by individuals whose beliefs are recursively shaped by the code itself (Levitt & March, 1988; March, 1991). Subsequent extensions introduce technology and ML as additional sources of information, prediction, or selection, but they continue to assume that technological outputs enter organizational learning through human-mediated codification or fixed model structures (Hendriks et al., 2024; Sturm et al., 2021). Thus, prior work treats technology as contributing to organizational learning without granting it autonomous participation in the generation, revision, and enactment of organizational beliefs.

The deployment of agentic AI in organizations challenges the assumptions underlying prior work. Unlike prior AI systems, agentic AI can generate sub-goals, evaluate its own performance, revise its learning strategy, and take consequential actions that modify organizational routines (Sapkota et al., 2025; Schneider, 2025; Shinn et al., 2023). These capabilities alter the architecture of organizational learning in

two ways. First, adaptive self-regulation makes the agent’s capability partly endogenous: learning no longer depends solely on human involvement, but also on the agent’s own evaluations of success and failure. Second, consequential action-taking creates a direct machine-mediated pathway from belief generation to routine enactment, reducing the extent to which organizational codification depends on human interpretation, endorsement, or intervention. Agentic AI therefore does not merely add another learner to March’s model; it changes the pathway through which beliefs become organizationally consequential.

These changes have theoretically ambiguous implications. On the one hand, agentic AI may enhance organizational learning by expanding the range of beliefs considered, accelerating search, and injecting machine-generated perspectives at scale. On the other hand, it may reduce belief diversity by accelerating convergence around self-generated beliefs, displacing human exploration, or entrenching erroneous routines before humans can detect or correct them. The central theoretical problem is therefore not whether agentic AI improves or harms organizational learning, but when its autonomy enhances knowledge accumulation and when it erodes the belief diversity on which long-run adaptation depends. We address this problem by extending March’s (1991) framework with a new class of learner-actor whose constitutive features—goal-directed reasoning, adaptive self-regulation, and consequential action-taking—are explicitly mapped onto the dynamics of organizational knowledge accumulation (level) and belief diversity.

Research Method

To address the research questions, in this study we opted for ABM and simulations following simulation-based research guidelines (Davis et al., 2007; Sargent, 2010). Agentic AI remains too recent a phenomenon to generate longitudinal data at the temporal scale necessary to observe organizational learning dynamics. Moreover, its effects arise from complex, co-evolving interactions between human and artificial agents that are not readily captured by closed-form analytical approaches. The nature of the phenomenon further constrains empirical observation, as key mechanisms unfold within adaptive systems that are difficult to isolate and examine in field settings.

In this context, ABM emerges as an appropriate methodological approach (Haki et al., 2020, 2024). It is a canonical tool in the organizational learning tradition (Fang et al., 2010; March, 1991; Miller et al., 2006), has been extended to model human-machine learning interactions (Sturm et al., 2021), and enables researchers to explicitly encode theoretical mechanisms and systematically examine their implications through controlled parameter variation (Nan, 2011). Following this tradition, we develop an agent-based model that builds on and extends March’s (1991) mutual-learning framework by introducing agentic AI as a distinct class of learner, alongside human agents and the organizational code. We further introduce three new model elements, each grounded in one of the three constitutive features of agentic AI outlined in the previous section, together with a set of inter-agent learning channels that become theoretically consequential once human and artificial learners coexist. Prior ABMs have extended March’s (1991) framework with passive ML agents (e.g., Sturm et al., 2021). We position our model contribution not as an incremental extension of those ML-based models but as a parallel extension of March’s framework in which the artificial agent is itself agentic—capable of reasoning across multiple dimensions, regulating its own learning, and acting directly on the organization.

Baseline Model: March’s (1991) Mutual-Learning Framework

Reality, agents, and beliefs. Building on March (1991), reality is represented as an m -dimensional vector whose entries take values in $\{-1, +1\}$. These values of reality are two different states per dimension that are unknown to the organization and its members. The organization consists of n human agents and an organizational code. Human beliefs and the code are represented by m -dimensional belief vectors with entries in $\{-1, 0, +1\}$, where 0 denotes the absence of a belief. The percentage of belief entries that correctly match reality defines a human agent’s or the organizational code’s knowledge level. Initial human beliefs are drawn uniformly from $\{-1, 0, +1\}$, the organizational code is initialized at 0 on all dimensions, and reality is drawn uniformly from $\{-1, +1\}$ on each dimension.

Mutual learning. Building on March (1991), in each period, humans interact with the organizational code in different ways: for each dimension of a human’s belief vector, the belief changes to the code’s value with probability p_1 whenever the code’s belief on that dimension is non-zero. Simultaneously, for each

dimension, the majority belief of knowledgeable humans, which we define as those whose beliefs match reality on more dimensions than the code does, is codified into the organizational code with probability p_2 .

Variety, environmental turbulence, and boundary openness. Inspired by Strum et.al (2021), our model includes two additional mechanisms to the March's (1991) baseline model, which are essential to make the model more realistic. (i) External stimulation: with probability p_{Ext} , a human adopts a fresh random belief (± 1 by coin flip), representing exploration from outside the organization's current code. (ii) Environmental turbulence: in each period, each reality dimension flips its sign with probability p_{Turb} , representing a non-stationary environment in which organizational knowledge may become outdated. Following March (1991), we conceptualize environmental turbulence as the exogenous rate at which the reality to which organizational beliefs must correspond changes, operationalized as the per-period probability that each reality dimension inverts (p_{Turb}). Theoretically, what distinguishes a turbulent from a stable environment is not the absolute rate of change but its relation to the organization's capacity to update beliefs: turbulence governs how quickly accumulated knowledge becomes obsolete, so its effects depend on whether environmental change outpaces the organization's learning processes (Kane & Alavi, 2007; Posen & Levinthal, 2012).

Extending the Baseline: Three Agentic-AI Features

To adapt March's (1991) framework to the domain of agentic AI, we extend the model by conceptualizing AI entities as organizational members, hereafter referred to as *artificial agents*. This integration is operationalized through three distinct model extensions (E1–E3, see Table 1).

E1 — Scope of autonomous reasoning (grounded in feature 1). In addition to its n human agents, an organization uses j artificial agents. Each artificial agent i covers k_i distinct dimensions of the m -dimensional reality vector, with $1 \leq k_i \leq m$. The artificial agent's belief vector holds 0 outside its scope and values in $\{-1, +1\}$ on its k_i in-scope dimensions. Scopes may overlap across artificial agents.

The parameter k_i operationalizes Feature 1—autonomous goal-directed reasoning (Sapkota et al., 2025; Schneider, 2025)—by indexing the breadth of the artificial agent's deliberative scope. The k_i parameter in our model captures this capacity as the number of reality dimensions over which a single artificial agent can form beliefs. Overlapping scopes reflect real-world deployments in which multiple artificial agents reason about related aspects of the same domain.

E2 — Self-regulated learning capability (grounded in feature 2). According to feature 2, each artificial agent can have a learning capability. Besides this, humans can decide to reconfigure artificial agents when it is necessary. Human reconfiguration may occur due to the overfitting of the artificial agent or changes in the tasks assigned to artificial agents. In practice, this reconfiguration could happen through changing the artificial agent architecture or its orchestration (Sapkota et al., 2025). In each period t , on each in-scope dimension, an artificial agent's belief becomes correct with probability q_t and incorrect with probability $1 - q_t$. The artificial agent updates q_t endogenously through self-regulation. Let t_{eff} denote the number of periods elapsed since the artificial agent was last reconfigured. Then:

$$q_t = 1 / (1 + \exp(-\sigma \cdot t_{eff} + \beta + \varepsilon_t)), \quad \text{with } \beta = \ln((1 - q_{Init}) / q_{Init}) \text{ and } \varepsilon_t \sim N(0, \delta^2).$$

In each period, each artificial agent is reconfigured by human intervention with probability $p_{Reconfig}$, upon which t_{eff} is reset to zero (so that q_t returns to initial learning capability (denotes as q_{Init})). The parameter σ governs the intensity of artificial agent's self-regulation. As the self-regulation cannot always be monotonically increasing due to reasons like AI hallucination, we introduced a random noise to self-regulation function, which its magnitude is set by δ . The logistic form is a standard way of representing learning dynamics with diminishing returns and has been used in prior ABM extensions of March (e.g., Sturm et al., (2021)). What is new in E2 is that the monotone argument of the logistic is the artificial agent's own accumulated experience t_{eff} rather than a function of human agreement, so the agent can adjust its own learning without human intervention. Critically, σ is bounded and imperfect: the artificial agent may optimize for proxy signals rather than true fitness and may drift locally-optimally but globally-suboptimal.

E3 — Consequential action taking as a direct path to influence organizational code (grounded in feature 3). In each period, for each dimension, an artificial agent's belief influences organizational knowledge through consequential action: with probability p_{Act} , the agent acts directly on its

in-scope dimension of the organizational code, overwriting the code's value with its own belief—bypassing the deliberative, human-mediated buffer. Per dimension, the channel is selected by a lottery with weights that we hold at (0.5, 0.5). This element operationalizes Feature 3—consequential action-taking—and is where our model departs most sharply from March' (1991). In March's original formulation, only human beliefs influence the organizational code, and then only through the deliberative codification process governed by p_2 . Feature 3 introduces a direct channel that bypasses codification entirely: when the artificial agent acts, it modifies the organizational substrate that other agents (human and artificial) subsequently encounter. The parameter p_{Act} reflects the scope of delegated autonomy granted to the artificial agent (Floridi, 2025). A low p_{Act} corresponds to tight governance and supervisory delegation (Baird & Maruping, 2021), while a high p_{Act} corresponds to broad delegation and generates the agency costs theorized by Humberd and Latham (2026).

Inter-Agent Learning Channels

March's (1991) original framework abstracts from direct inter-agent learning: humans learn from the code and the code learns from humans, but there is no channel for a human to learn from a peer, for an artificial agent to learn from a human, or for a human to learn from an artificial agent. Once artificial agents share the organizational setting with humans, these abstractions cease to be innocuous. Agentic systems increasingly learn from and influence human colleagues in situ through joint problem-solving, tool use, and direct delegation. To reflect this, we introduce three inter-agent learning channels with per-period, per-dimension activation probabilities: p_{HH} (human $\leftrightarrow$ human learning), p_{HA} (AI $\rightarrow$ human direct learning), and p_{AH} (human $\rightarrow$ AI direct learning). These channels sit alongside, rather than replacing, the organizational code-mediated channels p_1 and p_2 .

Nested baselines. The extensions E1–E3 together with the inter-agent channels nest two meaningful baselines. First, setting all agentic-AI parameters to their neutral values ($j = 0$, or equivalently $k_i = 0$) and $p_{HH} = p_{HA} = p_{AH} = 0$ recovers March's (1991) human-only model. Second, setting $k_i = 1$, $\sigma = 0$, $p_{Act} = 0$, and $p_{HH} = p_{HA} = p_{AH} = 0$ yields passive, single-dimension artificial learners whose only influence on the code is mediated codification, matching the design of prior ML extensions (Sturm et al., 2021).

Dependent and Manipulated Variables

To address the research questions and in line with the constitutive features of agentic AI, the simulation model examines the effects of a set of manipulated variables on a set of dependent variables. Specifically, the model manipulates the following variables: (1) *environmental turbulence* (p_{Turb}), (2) the rate of *consequential action-taking* by artificial agents (p_{Act}), (3) the intensity of *human reconfiguration* of artificial agents ($p_{Reconfig}$), and (4) the intensity of *self-regulation* (σ). The effects of these variables are evaluated on two dependent variables: (1) *organizational knowledge level*, representing the extent of knowledge accumulation in the organization, measured as the percentage of code dimensions that match reality dimensions; and (2) *belief diversity*, measured as the average pairwise belief distance across all agents (human and artificial) and all dimensions.

Baseline (March 1991) Assumptions	Our Extensions	Variables
Only human agents hold beliefs, across all m dimensions	Extension 1 introduces artificial agents, which have beliefs over more than one dimension and less than all.	$0 < k_i < m$
No artificial learners; humans update only via code socialization, peers, and exploration	Extension 2 enables artificial agents to self-regulate during the simulation and provide human agents a control through reconfiguration.	Reconfiguration intensity of artificial agents by human ($p_{Reconfig}$) and intensity of self-regulation (σ)
Organizational code updated only through	Extension 3 enables artificial agents to perform consequential action taking,	Rate of consequential action taking by artificial agents (p_{Act})

Baseline (March 1991) Assumptions	Our Extensions	Variables
human codification at rate p_2	which directly can influence organizational code.	
Table 1. Extensions of March's (1991) Model		

Model Operation and Simulation Setup

Each simulation run comprises T periods. Within every period, the model executes five steps in a fixed order. The ordering reflects the temporal dependencies of the underlying theory: environment precedes agents, individual learning precedes organizational codification, and outcome measurement is the last (see Figure 1).

Step 1 — Environmental turbulence. For each reality dimension, a Bernoulli (p_{Turb}) draw determines whether that dimension's sign flips.

Step 2 — Human belief update. For each human agent and each dimension, a weighted lottery determines the learning channel, selecting among code socialization (p_1), external exploration (p_{Ext}), human–human peer learning (p_{HH}), or AI→human direct learning (p_{HA}), each with its corresponding activation probability. Channel weights are uniform by default (0.25 each).

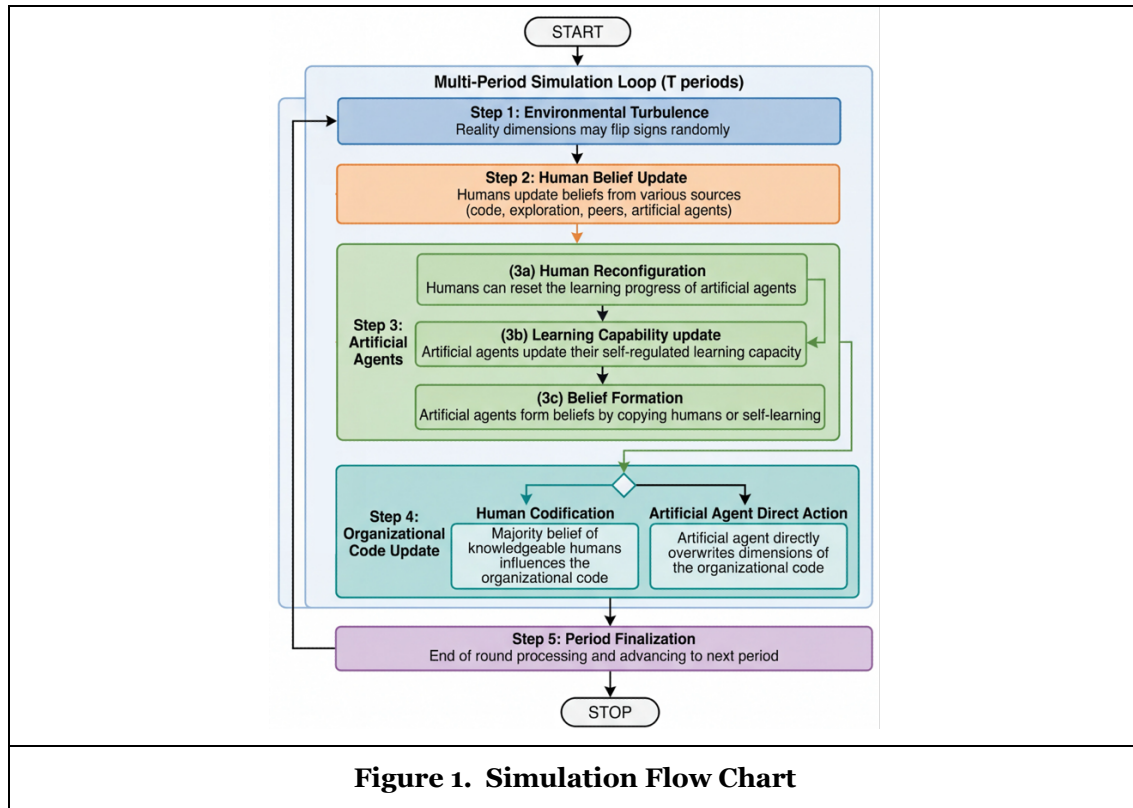


Figure 1. Simulation Flow Chart

Step 3 — Artificial agent cognition. For each artificial agent, the step decomposes into three sub-steps:

- (3a) Reconfiguration: with probability p_{Reconfig} the agent is reset ($t_{\text{eff}} \rightarrow 0$), otherwise t_{eff} is incremented.
- (3b) q_t update: the logistic self-regulation equation from E2 is evaluated with a fresh Gaussian perturbation ε_t .

- (3c) Belief formation: for each in-scope dimension, a Bernoulli (p_{AH}) draw selects between copying a sampled human's belief and self-learning (correct with probability q_i). The ordering of steps 2 and 3 ensures that humans observed by the AI in step 3c are those carried into the period, preventing within-period circularity.

Step 4 — Organizational code update. For each dimension, a binary lottery selects between two mechanisms: (i) human codification, whereby the majority belief among knowledgeable humans is codified according to March (1991) acceptance probability $1 - (1 - p_2)^k$, where k denotes the majority count; and (ii) artificial agent direct action, with probability p_{Act} . At each period, every dimension covered by at least one artificial agent enters a Bernoulli lottery with probability p_{Act} . If triggered, the organizational code on that dimension is overwritten by the majority of the covering artificial agents' non-zero beliefs, with ties broken uniformly at random. The Bernoulli draw is performed once per dimension rather than once per artificial agent, so the per-dimension rate of artificial agent direct action is invariant to the degree of scope overlap among artificial agents (E3).

Step 5 — Finalizing each simulation round. After updating all agents' beliefs and their interaction with the organizational code, the model calculates dependent variables and moves to the next round of simulation.

We implement the model in Python using the Mesa framework. Each belief dimension of reality is initialized uniformly from $\{-1, +1\}$; human belief vectors are initialized uniformly from $\{-1, 0, +1\}$; and the organizational code and artificial agent belief vectors are initialized at 0 on all dimensions. Building on March (1991) and its subsequent extensions (Sturm et al., 2021), the default parameter values are $m = 30$, $n = 50$, $j = 10$, $p_1 = p_2 = 0.5$, $p_{Ext} = 0.1$, and the three inter-agent learning channels are set to $p_{HH} = p_{HA} = p_{AH} = 0.2$. The artificial agent parameters are fixed or varied as follows: $k_i = 5$, $\sigma \in [0, 1]$, $p_{Act} \in [0, 1]$, $p_{Reconfig} \in \{0, 0.01, 0.05, 0.1\}$, $\delta = 0.1$, $p_{Turb} \in \{0, 0.02, 0.2\}$, and $q_{Init} = 0.5$. Each parameter combination is run for 80 stochastic iterations of 200 periods each.

Results

In this section, we first examine the effects of the manipulated variables, namely the rate of *consequential action-taking* by artificial agents (p_{Act}), the intensity of *human reconfiguration* of artificial agents ($p_{Reconfig}$), and the intensity of *self-regulation* (σ), on the dependent variables, namely long-term organizational *knowledge level* and *belief diversity*. To account for interaction effects among these variables, we group them into three stylized artificial agency regimes: human oversight, AI autonomy, and balanced. The human oversight regime reflects high human involvement and is characterized by $\sigma = 0.1$, $p_{Act} = 0.1$, $p_{Reconfig} = 0.1$ configuration. Low values for p_{Act} and σ indicate that artificial agents possess limited consequential action-taking capability and learning capabilities; when coupled with a high $p_{Reconfig}$, these parameters ensure a greater degree of human control over artificial agents. In contrast, the AI autonomy regime reflects high autonomy of artificial agents, characterized by $\sigma = 0.8$, $p_{Act} = 0.7$, $p_{Reconfig} = 0.01$ configuration. The balanced regime lies between these two extremes ($\sigma = 0.5$, $p_{Act} = 0.3$, $p_{Reconfig} = 0.05$).

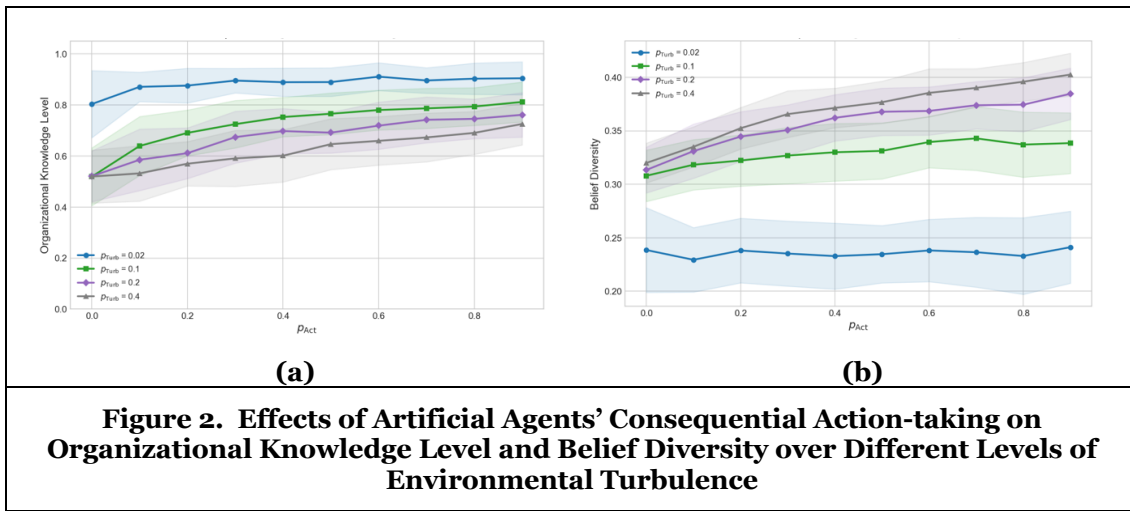
To assess the individual and interaction effects of the manipulated variables, we distinguish varying levels of *environmental turbulence* (p_{Turb}), ranging from stable to highly turbulent conditions. A stable environment is defined by a low rate of change (2%, $p_{Turb} = 0.02$), whereas a turbulent environment is defined by a high rate of change (40%, $p_{Turb} = 0.4$). The ranges for p_{Turb} and $p_{Reconfig}$ are in lined with existing models (Sturm et al., 2021).

Effects of Consequential Action-Taking on Organizational Knowledge Level and Belief Diversity

Figure 2 shows that the effect of consequential action-taking is highly dependent on the level of environmental turbulence. In stable environments ($p_{Turb} = 0.02$), its influence is minimal. Across the entire range of consequential action-taking, knowledge level remains consistently high, close to its upper bound (≈ 0.88), while belief diversity remains low close to its lower bound (≈ 0.24). Therefore, when the environment is relatively predictable, increasing the degree of action-taking does not meaningfully change either what the organization knows or how diverse its knowledge base is. The organization appears to have already converged, leaving little room for further improvement through additional action.

This pattern changes substantially as the environment becomes turbulent. Under higher levels of environmental turbulence, consequential action-taking emerges as a central mechanism shaping both outcomes. As action-taking increases, knowledge level rises significantly, indicating improved alignment with the evolving environment. At the same time, belief diversity also increases, suggesting that the organization is able to sustain a broader range of perspectives while adapting. For instance, under high turbulence ($p_{\text{Turb}} = 0.4$), increasing action-taking leads to a substantial improvement in knowledge levels (from 0.52 at $p_{\text{Act}} = 0$ to 0.73 at $p_{\text{Act}} = 0.9$), alongside a noticeable expansion in belief diversity (from 0.32 to 0.40). This simultaneous increase is particularly important, as it suggests that organizations are not forced to trade off accuracy for diversity under these conditions.

Therefore, consequential action-taking functions as a turbulence-contingent lever. Its value does not lie in universally improving outcomes, but in enabling adaptation when the environment is unstable. In turbulent settings, consequential action-taking allows organizations to respond more quickly to shocks, update their knowledge more effectively, and maintain sufficient variation to cope with ongoing change. In contrast, in stable environments, where both knowledge and diversity have largely stabilized, additional action-taking provides little incremental benefit.



Effects of Human Reconfiguration on Organizational Knowledge Level and Belief Diversity

Figure 3 shows that the effect of human reconfiguration of artificial agents on outcomes is generally modest and varies depending on the level of environmental turbulence. Across all environmental conditions, its influence on knowledge level is limited and slightly negative (a 0.02–0.05 reduction as p_{Reconfig} moves from 0 to 0.1). As human reconfiguration increases, knowledge level declines marginally, suggesting that reconfiguring agents does not contribute to improving the organization's alignment with the environment and may even introduce small inefficiencies.

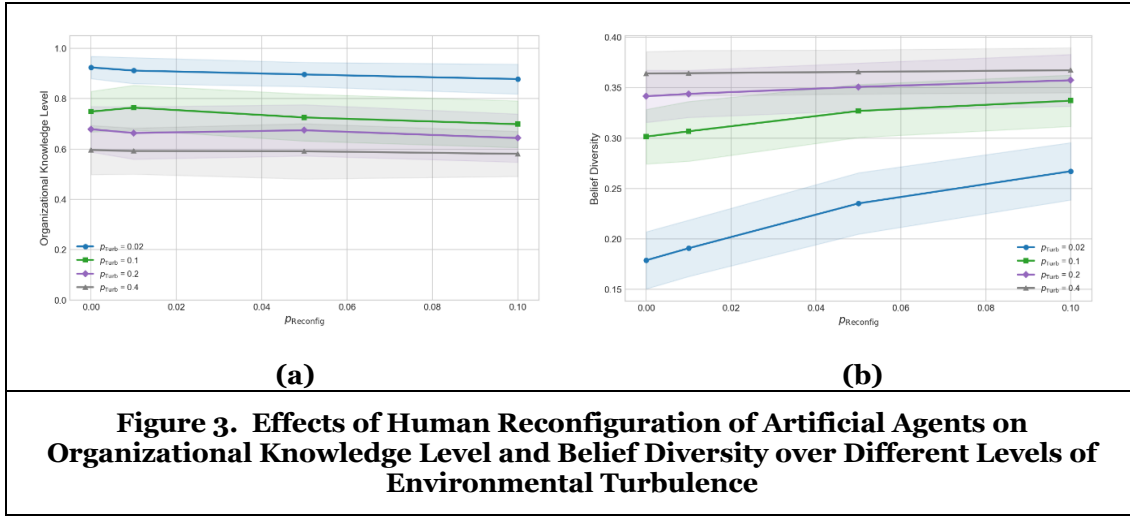
The effect on belief diversity, however, is more nuanced and clearly shaped by environmental turbulence. In stable environments ($p_{\text{Turb}} = 0.02$), increasing human reconfiguration leads to a noticeable rise in belief diversity, from 0.18 to 0.27. This indicates that human reconfiguration helps introduce variation into the system, counteracting the tendency toward premature convergence when the environment itself does not provide sufficient change. In such settings, human reconfiguration plays a constructive role by maintaining a broader set of perspectives within the organization.

This pattern does not hold under turbulent conditions. When environmental turbulence is high ($p_{\text{Turb}} = 0.4$), belief diversity remains largely constant across the full range of human reconfiguration. In this case, the environment itself continuously generates variation, reducing the need for additional diversity introduced through human reconfiguration. As a result, the marginal contribution of human reconfiguration becomes negligible.

Therefore, human reconfiguration of artificial agents functions as a supplementary diversity mechanism whose effectiveness depends on the surrounding environment. In stable settings, it helps sustain diversity by injecting variation into an otherwise converging system. In contrast, under turbulent conditions, where diversity is already maintained through ongoing environmental change, its role becomes largely redundant.

Effects of Self-Regulation on Organizational Knowledge Level and Belief Diversity

Figure 4 reveals the most pronounced trade-off in our parameter space, centered on the role of self-regulation. Even at low levels, self-regulation leads to a sharp increase in knowledge level across all levels of environmental turbulence (5–13 percentage points between $\sigma = 0$ and $\sigma = 0.1$). Moving from no self-regulation to minimal self-regulation produces a noticeable improvement, after which the gains quickly plateau. This suggests that a small degree of self-regulation is sufficient to substantially improve the organization's alignment with its environment, while higher levels provide only limited additional benefit.



These gains in knowledge level, however, come at the expense of belief diversity, particularly in stable environments. Under low turbulence ($p_{\text{Turb}} = 0.02$), increasing self-regulation leads to a substantial reduction in diversity, with values dropping from 0.37 at no self-regulation to 0.20 at high levels. This indicates that self-regulating agents converge more quickly toward a shared understanding, reducing variation in the organization's knowledge base. In such settings, self-regulation accelerates convergence on an existing reality, resulting in more homogeneous beliefs.

In contrast, this trade-off becomes much less pronounced under turbulent conditions. When environmental turbulence is high ($p_{\text{Turb}} = 0.4$), belief diversity remains largely stable across different levels of self-regulation. The continuous introduction of change prevents the organization from fully converging, even when agents are highly self-regulating. As a result, the reduction in diversity typically associated with self-regulation is effectively offset by ongoing environmental variation.

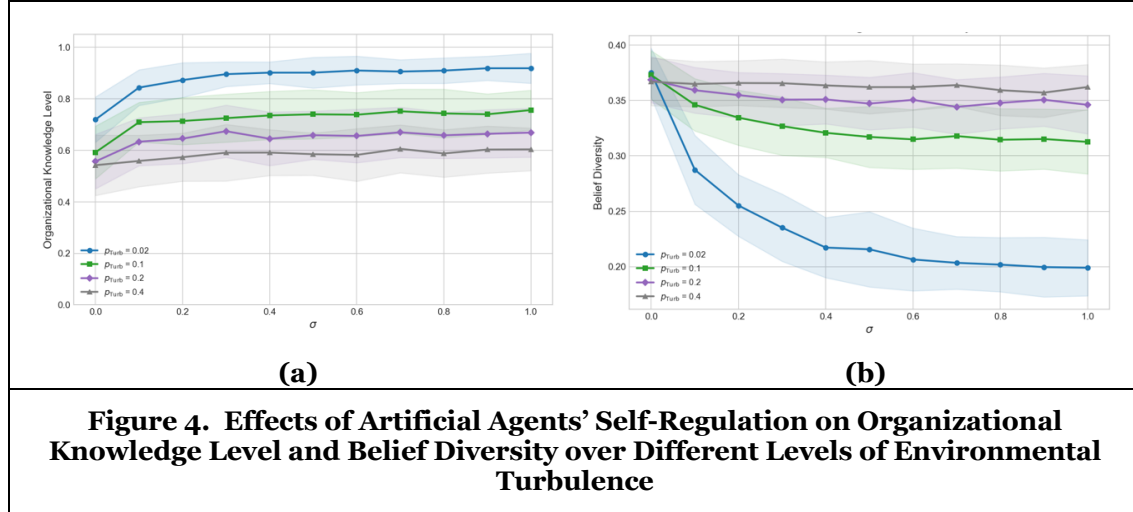
Therefore, self-regulation introduces a fundamental trade-off between knowledge level and belief diversity, the strength of which depends on the environment. In stable settings, self-regulation improves accuracy but drives convergence and homogenization. In turbulent settings, however, the environment itself limits convergence, allowing organizations to benefit from improved knowledge without sacrificing diversity. This highlights that the implications of self-regulation cannot be assessed in isolation but must be understood in relation to the dynamics of the environment in which it operates.

Organizational Knowledge Level and Belief Diversity Across Three Artificial Agency Regimes

To better understand how these mechanisms interact, we define three stylized artificial agency configuration regimes, namely human oversight, AI autonomy, and balanced, and examine their

performance at the two endpoints of the turbulence spectrum. Figure 5 compares these configurations in terms of knowledge level and belief diversity under stable and turbulent conditions.

In terms of knowledge level, a consistent pattern emerges across environments. The AI autonomy configuration outperforms the others in both stable and turbulent settings, achieving the highest levels of knowledge. Under stable conditions, knowledge levels are close to their upper bound across all configurations, but AI autonomy still maintains a noticeable advantage over human oversight.



This ordering persists under turbulence, although overall knowledge levels decline due to the increased difficulty of tracking a changing environment. Even then, AI autonomy retains a clear lead, suggesting that higher levels of self-regulation and action-taking enable more effective alignment with environmental changes.

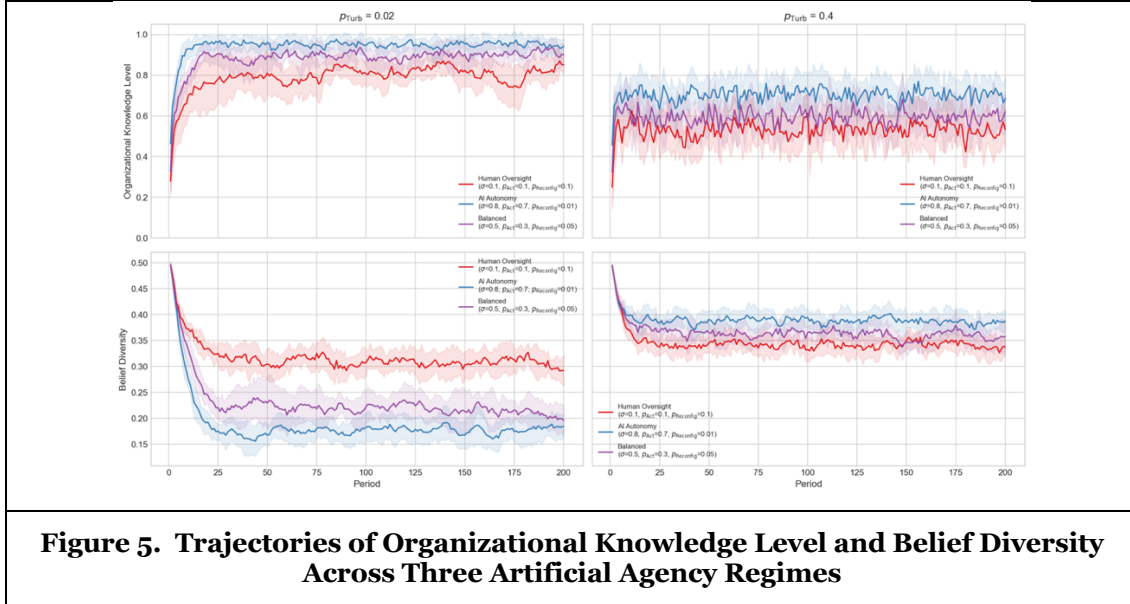
The pattern for belief diversity, however, depends strongly on the level of environmental turbulence. In stable environments, human oversight yields the highest diversity, while AI autonomy results in the lowest. This indicates that configurations emphasizing human involvement help maintain variation when the environment does not naturally generate it. In contrast, AI autonomy accelerates convergence, leading to more homogeneous knowledge structures.

Under turbulent conditions, this ordering reverses. AI autonomy now produces the highest levels of diversity, while human oversight yields the lowest. In this case, the same mechanisms that previously drove convergence instead allow the organization to track multiple evolving dimensions of the environment more effectively. Higher rate of consequential action-taking by artificial agents and intensity of self-regulation enable agents to respond to continuous change, sustaining a broader range of knowledge. Human oversight, by comparison, appears less able to keep pace with environmental dynamics, resulting in lower diversity.

Therefore, the performance of agentic AI governance configurations is not fixed but depends on environmental conditions. The same configuration that leads to homogenization in stable environments can sustain high diversity under turbulence. This shift occurs because environmental change continuously introduces new variation, which can be captured and processed more effectively by configurations that emphasize autonomy and rapid action. As a result, the relative advantages of human oversight and AI autonomy are context-dependent, underscoring the importance of aligning governance configurations with the dynamics of the environment.

Discussion

This study sets out to examine how agentic AI, characterized by autonomous goal-directed reasoning, self-regulation, and consequential action-taking, reshapes the canonical dynamics of organizational learning. Building on March's (1991) mutual-learning model, we introduced a new class of learners whose behavior is not mediated exclusively through human interpretation but can directly alter organizational routines.



In doing so, we depart from prior IS research that conceptualizes AI either as a tool (Humberd & Latham, 2026) or as a passive learner embedded in human-controlled processes (Sturm et al., 2021) and instead model AI as a delegated organizational actor (Baird & Maruping, 2021; Floridi, 2025).

Implications for Research and Practice

Across our simulations, three findings emerge that collectively challenge current assumptions of organizational learning theory. First, the value of consequential action-taking is sharply contingent on environmental turbulence, extending Posen & Levinthal (2012). It is largely redundant in stable environments but becomes essential when the rate of environmental turbulence exceeds the organization's ability to update beliefs through deliberation, where it enables faster adaptation without forcing a trade-off between accuracy and diversity. This finding reframes the role of organizational routines. Rather than being solely repositories of accumulated knowledge, routines become objects of continuous intervention by artificial agents. The organizational code, in turn, is no longer merely an aggregator of beliefs but an evolving substrate shaped by both human cognition and machine action.

Second, human reconfiguration, which is central in prior AI-based models (Sturm et al., 2021), loses its primacy once learning capability becomes endogenous. It acts as a supplementary mechanism to sustain diversity in stable environments but becomes largely redundant in turbulent settings where environmental change itself generates sufficient variation. Our model also shows that once artificial agents possess self-regulation capabilities, learning becomes endogenous. This shift has two main implications. It alters the locus of adaptation. Instead of residing primarily in human reconfiguration decisions, adaptive capacity migrates into the artificial agent itself, consistent with Floridi's (2025) concept of artificial agency. In addition, it changes the function of human oversight. Rather than improving AI accuracy, human intervention primarily serves to preserve epistemic heterogeneity by interrupting convergence processes.

Third, agentic AI introduces a knowledge-diversity trade-off whose direction is not fixed but flips with environmental conditions, complicating the extant understanding of exploration-exploitation dichotomy (March, 1991). In stable environments, agentic AI may intensify exploitation by accelerating convergence around effective beliefs, whereas in turbulent environments the same capabilities may preserve or regenerate diversity by continuously revising beliefs in response to shifting realities. This finding extends Posen and Levinthal (2012) by demonstrating that exploration can be endogenously sustained not only by deliberate search but also by the interaction between autonomous agents and a shifting environment. In this sense, exploration becomes an emergent property of the system rather than a deliberate strategic choice. Our findings can be read as identifying boundary conditions for two influential accounts of machine involvement in organizational learning. Balasubramanian et al. (2022) proposed that substituting human decision-making with ML reduces routine diversity and exacerbates learning myopia, because ML selects routine variants on the basis of aggregated historical data. Our results reproduce this concern under stable

conditions: self-regulating, high-autonomy agents drive belief diversity toward its floor, the formal analogue of the homogenization they theorize. Under environmental turbulence, however, the prediction reverses — autonomy sustains the highest diversity in our simulations. The reversal follows from a mechanism difference that their propositions did not need to consider: selection-based ML is inherently variance-reducing because it chooses among existing variants, whereas an agent that generates and enacts its own beliefs becomes variance-*generating* once the environment continually supplies novelty to track.

In sum, these results suggests that agentic AI does not simply “improve” organizational learning; it reconfigures the mechanisms through which learning unfolds. Our study also contributes to the broader IS discourse on AI by providing a formal representation of agentic AI as an organizational actor. While prior conceptual work has argued for this shift (Berente et al., 2021; Lyytinen et al., 2021), empirical and formal treatments have lagged behind. By embedding agentic AI within a well-established theoretical framework, we demonstrate that its defining features (autonomous goal-directed reasoning, self-regulation, and consequential action-taking) are not merely descriptive attributes but mechanisms with distinct and interacting effects on organizational learning outcomes and their trade-off. This moves the literature beyond categorical claims about AI agency toward a more precise understanding of how different dimensions of agency shape organizational dynamics.

As a managerial implication, our findings suggest that organizations should move away from uniform governance approaches to agentic AI and toward contingent, environment-sensitive configurations in the context of organizational learning. In stable environments, limiting consequential action-taking and maintaining human reconfiguration can help preserve diversity and avoid premature convergence. In turbulent environments, higher levels of autonomy and self-regulation can enhance both knowledge level and diversity by enabling faster adaptation. This implies that the key managerial capability is not simply deciding how much to delegate to AI, but developing the ability to continuously recalibrate delegation in response to changing environmental conditions.

Boundary Conditions and Future Research

Several boundary conditions should be considered when interpreting our insights. First, while most model choices and values are informed by existing models, our parameters are not empirically calibrated. Future work is called to link model parameters, such as self-regulation and action-taking, to observable characteristics of real-world AI systems. Although our model has the capability to be reduced to March’s original model which can serve as form of validation, computer science literature (i.e., Shinn et al., 2023) shown empirically the benefits of features like self-regulation in agentic AI. Second, we assume homogeneous artificial agents, whereas organizations usually deploy heterogeneous AI systems with varying capabilities and governance structures. Building on our model, incorporating heterogeneous artificial agents could generate additional and potentially novel insights. Third, our model abstracts from goal misalignment, a central concern in principal–agent theory (Eisenhardt, 1989). Allowing agents to pursue objectives that diverge from organizational goals would likely introduce new dynamics related to trust, control, and strategic behavior. Fourth, our measure of belief diversity captures dispersion but not its quality. Distinguishing between productive diversity and noise remains an important empirical challenge. Fifth, the propositions related to AI governance is only limited to the context of organizational learning and can not be generalized to all the influence of Agentic AI in an organizations. Finally, future research is encouraged to connect these learning dynamics to broader organizational outcomes, such as innovation and performance, thereby integrating agentic AI into the broader theory of the firm.

References

- Ågerfalk, P. J. (2020). Artificial intelligence as digital agency. *European Journal of Information Systems*, 29(1), 1–8. <https://doi.org/10.1080/0960085X.2020.1721947>
- Alavi, M., & Leidner, D. E. (2001). Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues. *MIS Quarterly*, 25(1), 107–136. <https://doi.org/10.2307/3250961>
- Baird, A., & Maruping, L. M. (2021). The Next Generation of Research on Is Use: A Theoretical Framework of Delegation to and from Agentic Is Artifacts. *MIS Quarterly*, 45(1), 315–341. <https://doi.org/10.25300/MISQ/2021/15882>

- Balasubramanian, N., Ye, Y., & Xu, M. (2022). Substituting Human Decision-Making with Machine Learning: Implications for Organizational Learning. *Academy of Management Review*, 47(3), 448–465. <https://doi.org/10.5465/amr.2019.0470>
- Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, 45(3), 1433–1450. <https://doi.org/10.25300/MISQ/2021/16274>
- Choudhary, V., Marchetti, A., Shrestha, Y. R., & Puranam, P. (2025). Human-AI ensembles: When can they work? *Journal of Management*, 51(2), 536–569.
- Csaszar, F. A., & Levinthal, D. A. (2016). Mental representation and the discovery of new strategies. *Strategic Management Journal*, 37(10), 2031–2049. <https://doi.org/10.1002/smj.2440>
- Davis, J. P., Eisenhardt, K. M., & Bingham, C. B. (2007). Developing Theory Through Simulation Methods. *Academy of Management Review*, 32(2), 480–499.
- Denrell, J., & March, J. G. (2001). Adaptation as information restriction: The hot stove effect. *Organization Science*, 12(5), 523–538. <https://doi.org/10.1287/orsc.12.5.523.10092>
- Eisenhardt, K. M. (1989). Agency theory: An assessment and review. *Academy of Management Review*, 14(1), 57–74.
- Fang, C., Lee, J., & Schilling, M. A. (2010). Balancing exploration and exploitation through structural design: The isolation of subgroups and organizational learning. *Organization Science*, 21(3), 625–642. <https://doi.org/10.1287/orsc.1090.0468>
- Faraj, S., Pachidi, S., & Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. *Information and Organization*, 28(1), 62–70.
- Floridi, L. (2025). AI as Agency without Intelligence: On Artificial Intelligence as a New Form of Artificial Agency and the Multiple Realisability of Agency Thesis. *Philosophy & Technology*, 38(1), 30. <https://doi.org/10.1007/s13347-025-00858-9>
- Haki, K., Beese, J., Aier, S., & Winter, R. (2020). The evolution of information systems architecture: An agent-based simulation model. *MIS Quarterly*, 44(1), 155–184.
- Haki, K., Safaei, D., Magan, A., & Griffiths, M. (2025). Integrating generative AI into enterprise platforms: Insights from salesforce. *Information Systems Journal*, 35(5), 1497–1512.
- Haki, K., Tanriverdi, H., Safaei, D., Schmid, M., Aier, S., & Winter, R. (2024). Generativity and profitability on B2B innovation platforms: A simulation-based theory development. *MIS Quarterly*, 48(2), 583–612.
- Hendriks, P., Sturm, T., Geis, M., Grimminger, T., & Mast, B. (2024). Exploring collaboration in human-artificial intelligence teams: A design science approach to team-AI collaboration systems. *PACIS 2024 Proceedings*.
- Humberd, B. K., & Latham, S. F. (2026). When AI Becomes an Agent of the Firm: Examining the Evolution of AI in Organizations Through an Agency Theory Lens. *Journal of Management Studies*, 63(2), 668–694. <https://doi.org/10.1111/joms.13274>
- Iyengar, K., Sweeney, J. R., & Montealegre, R. (2015). Information Technology Use as a Learning Mechanism: The Impact of IT Use on Knowledge Transfer Effectiveness, Absorptive Capacity, and Franchisee Performance. *MIS Quarterly*, 39(3), 615–641. <https://doi.org/10.25300/MISQ/2015/39.3.05>
- Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255–260. <https://doi.org/10.1126/science.aaa8415>
- Kane, G. C., & Alavi, M. (2007). Information technology and organizational learning: An investigation of exploration and exploitation processes. *Organization Science*, 18(5), 796–812.
- Kasirzadeh, A., & Gabriel, I. (2025). Characterizing AI Agents for Alignment and Governance (arXiv:2504.21848). *arXiv*. <https://doi.org/10.48550/arXiv.2504.21848>
- Lavie, D., Stettner, U., & Tushman, M. L. (2010). Exploration and exploitation within and across organizations. *Academy of Management Annals*, 4(1), 109–155.
- Lebovitz, S., Levina, N., & Lifshitz-Assaf, H. (2021). Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What. *MIS Quarterly*, 45(3), 1501–1526. <https://doi.org/10.25300/MISQ/2021/16564>
- Levinthal, D. A., & March, J. G. (1993). The myopia of learning. *Strategic Management Journal*, 14(S2), 95–112.
- Levitt, B., & March, J. G. (1988). Organizational learning. *Annual Review of Sociology*, 14(1), 319–338.
- Liu, T., Giorgi, S., Aich, A., Lahnala, A., Curtis, B., Ungar, L., & Sedoc, J. (2025). The illusion of empathy: How AI chatbots shape conversation perception. *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial*

- Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25, 39, 14327–14335. <https://doi.org/10.1609/aaai.v39i13.33569>
- Lytyinen, K., Nickerson, J. V., & King, J. L. (2021). Metahuman systems = humans + machines that learn. *Journal of Information Technology*, 36(4), 427–445. <https://doi.org/10.1177/0268396220915917>
- March, J. G. (1991). Exploration and Exploitation in Organizational Learning. *Organization Science*, 2(1), 71–87.
- Miller, K. D., Zhao, M., & Calantone, R. J. (2006). Adding Interpersonal Learning and Tacit Knowledge to March's Exploration-Exploitation Model. *The Academy of Management Journal*, 49(4), 709–722.
- Murray, A., Rhymer, J., & Sirmon, D. G. (2021). Humans and Technology: Forms of Conjoined Agency in Organizations. *Academy of Management Review*, 46(3), 552–571. <https://doi.org/10.5465/amr.2019.0186>
- Nan, N. (2011). Capturing Bottom-Up Information Technology Use Processes: A Complex Adaptive Systems Model. *MIS Quarterly*, 35(2), 505–A7.
- Posen, H. E., & Levinthal, D. A. (2012). Chasing a moving target: Exploitation and exploration in dynamic environments. *Management Science*, 58(3), 587–601. <https://doi.org/10.1287/mnsc.1110.1420>
- Puranam, P., & Swamy, M. (2016). How initial representations shape coupled learning processes. *Organization Science*, 27(2), 323–335.
- Raisch, S., & Krakowski, S. (2021). Artificial Intelligence and Management: The Automation–Augmentation Paradox. *Academy of Management Review*, 46(1), 192–210. <https://doi.org/10.5465/amr.2018.0072>
- Sapkota, R., Roumeliotis, K. I., & Karkee, M. (2025). Ai agents vs. Agentic ai: A conceptual taxonomy, applications and challenges. *Information Fusion*, 103599.
- Sargent, R. G. (2010). Verification and validation of simulation models. *Proceedings of the 2010 Winter Simulation Conference*, 166–183.
- Schneider, J. (2025). Generative to Agentic AI: Survey, Conceptualization, and Challenges (arXiv:2504.18875). *arXiv*. <https://doi.org/10.48550/arXiv.2504.18875>
- Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., Lee, T., Mishkin, P., Eloundou, T., Hickey, A., & others. (2023). Practices for governing agentic AI systems. Research Paper, *OpenAI*.
- Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 8634–8652.
- Stelmaszak, M., Joshi, M., & Constantiou, I. (2026). Artificial intelligence as an organizing capability arising from human-algorithm relations. *Journal of Management Studies*, 63(2), 335–365.
- Strich, F., Mayer, A.-S., & Fiedler, M. (2021). What Do I Do in a World of Artificial Intelligence? Investigating the Impact of Substitutive Decision-Making AI Systems on Employees' Professional Role Identity. *Journal of the Association for Information Systems*, 22(2). <https://doi.org/10.17705/1jais.00663>
- Sturm, T., Gerlach, J., Pumplun, L., Mesbah, N., Peters, F., Tauchert, C., Nan, N., & Buxmann, P. (2021). Coordinating Human and Machine Learning for Effective Organization Learning. *MIS Quarterly*, 45(3), 1581–1602.
- Vanneste, B., & Puranam, P. (2024). Artificial Intelligence, Trust, and Perceptions of Agency (SSRN Scholarly Paper No. 3897704). *Social Science Research Network*. <https://doi.org/10.2139/ssrn.3897704>
- Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), 186345. <https://doi.org/10.1007/s11704-024-40231-1>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., & others. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.
- Writer. (2026, April 7). WRITER Survey Finds 60% of Companies Plan to Lay Off Employees Who Won't Adopt AI. *WRITER*. <https://writer.com/blog/enterprise-ai-adoption-survey-results-press-release/>
- Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., ... Gui, T. (2025). The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2), 121101. <https://doi.org/10.1007/s11432-024-4222-0>