THE AXIOLOGY OF AI AGENTS

How Human-AI Interaction Transforms Values from Espoused to Enacted

Vern L. Glaser
University of Alberta
vglaser@ualberta.ca

Jennifer Sloan
University College of London

Rodrigo Valadão
Fundação Getulio Vargas (FGV)

Evelyn Micelotta
University of Vermont

Accepted for publication at Academy of Management Review

We thank editor Keith Leavitt and the three anonymous reviewers for their developmental guidance throughout the review process, which substantially strengthened this article. We gratefully acknowledge the support of the Social Sciences and Humanities Research Council of Canada and the Alberta School of Business at the University of Alberta. This research forms part of the UK Research and Innovation project "Innovating Across Sectors" (MR/Y034430/1; PI: Angela Aristidou). The analyses and conclusions presented here are those of the authors, not of the funding body.

ABSTRACT

Organizational scholars have long recognized that espoused values diverge from enacted values. LLM-based AI agents—autonomous software systems capable of perceiving environments, making decisions, collaborating, and taking actions to achieve defined goals—create a novel form of this classic problem. Unlike traditional technologies where humans retain interpretive control, AI agents autonomously navigate value trade-offs, and enacted values may differ from those espoused. How does the interaction of humans and AI transform espoused organizational values? Drawing on assemblage theory, we trace values transformation across three stacked assemblages (the LLM assemblage, the AI agent assemblage, and the algorithmic routine assemblage) through three transformation moments. During configuration, inherited orientations become explicit parameters; during embedding, parameters are integrated into organizational routines; during performance, values are enacted through the interplay of algorithmic pathways and improvised responses. Recursive feedback connects enacted outcomes to subsequent configurations, producing ongoing transformation rather than stable alignment. Our framework extends scholarship on espoused and enacted values by specifying technologically mediated transformation mechanisms produced through the interaction of humans and AI. We also extend research on values work by showing how AI agents enact values without recognizing when those values no longer fit, leaving human improvisation to diagnose where value translation has faltered.

Organizations are increasingly adopting AI agents: autonomous software systems capable of perceiving environments, making decisions, collaborating, and taking actions to achieve defined goals (Humberd & Latham, 2025; Sapkota et al., 2025; see also Wooldridge & Jennings, 1995). Advances in large language models (LLMs)[1] have accelerated this turn, enabling these agents to move past straightforward execution of predefined rules toward interpreting context and taking independent action (Luo et al., 2025). Such autonomy creates a distinctive organizational challenge. Organizations design AI agents to represent particular values (Følstad et al., 2018), but the agents must autonomously prioritize competing concerns such as responsiveness versus accuracy, efficiency versus fairness, and automation versus human oversight.

The gap between organizations' espoused values and enacted values is a long-recognized organizational problem (Argyris & Schön, 1974; Howell et al., 2012; Schein, 2010), one that AI agents now recast in novel, more challenging, form. The Air Canada chatbot case illustrates these higher stakes. When the company's AI agent prioritized efficiency over policy accuracy by incorrectly promising refunds, the resulting legal liability exposed how the values espoused in the system configuration diverged sharply from the values the AI agent enacted in practice (*Moffatt v. Air Canada*, 2024). Unlike backend algorithms, where such divergence may remain invisible, AI agents interact directly with clients, employees, and citizens, making the gap between espoused and enacted values immediately observable and consequential (Wallach & Allen, 2009). This distinctive positioning of AI agents as both bearers of organizational values

[1] Large language models (LLMs) are artificial intelligence systems based on transformer neural network architectures, trained on extensive text corpora to predict and generate human-like language. Through this training process, LLMs develop emergent capabilities such as reasoning, summarization, and contextual interpretation that arise from scale rather than explicit programming (Zhao et al., 2023). Examples include OpenAI's GPT-4, Anthropic's Claude, and Meta's Llama. For an accessible introduction, see Stanford University's "AI Demystified: Introduction to Large Language Models" (https://uit.stanford.edu/service/techtraining/ai-demystified/llm); for organizational applications and the AI technology stack, see Hosanagar & Krishnan (2024); for technical foundations, see Bommasani et al. (2022).

and autonomous decision-makers exposes a fundamental theoretical puzzle: *How does the interaction of humans and AI transform espoused organizational values?*

Two streams of scholarship address pieces of this question. Technical AI research treats values as parameters to be specified and optimized, translating abstract normative commitments into computational form. Approaches such as values-sensitive design integrate moral considerations with empirical investigation through iterative, stakeholder-centered processes (Friedman & Hendry, 2019; Shneiderman, 2022), and other frameworks emphasize participatory design, responsible innovation, or ethics-by-design principles (e.g., Stilgoe et al., 2013). At the algorithmic level, training-time alignment techniques like constitutional AI and reinforcement learning from human feedback (RLHF) attempt to translate human values into learnable parameters (Bai et al., 2022; Christiano et al., 2017), while operational safeguards such as prompt-layer guardrails filter outputs during use (Dong et al., 2024). When tensions arise between accuracy and efficiency or fairness and utility, this perspective proposes systematic solutions such as multi-objective optimization (La Cava, 2023), bias mitigation techniques (Mehrabi et al., 2022), or governance interventions spanning software engineering practices, organizational safety cultures, and independent oversight (Rahwan, 2018; Shneiderman, 2022). These approaches show how values are espoused through technical parameters during configuration, but they focus on the initial moment of espousal, leaving open what happens when AI agents subsequently encounter organizational life (Teodorescu et al., 2021).

Management research offers a distinct perspective that shifts focus from encoding values during the design process to the ongoing enactment of values in organizational life. Scholars have long understood values as enduring beliefs that guide action (Kraatz et al., 2020; Kraatz & Flores, 2015; Rokeach, 1973; Schwartz, 1992; Selznick, 1957), and contemporary theory shows

how they emerge through sociotechnical configurations in which human and material elements are deeply entangled (Gehman et al., 2013). Although designers may inscribe particular values into AI systems through technical choices (Akrich, 1992; Glaser, 2017; Monteiro, 2022), these inscriptions never fully determine outcomes; instead, values take shape in use through the dynamic interplay between these technical choices and evolving organizational practices (Orlikowski & Scott, 2008). In this process, human and technological elements enact conjoined agency (Murray et al., 2021), navigating tensions between automation and augmentation (Raisch & Krakowski, 2021; Raisch & Fomina, 2025) in ways that continually reshape how values are prioritized. This view foregrounds values work, the ongoing process whereby organizational members articulate, negotiate, and perform what is considered appropriate (Gehman et al., 2013), and helps explain why espoused values frequently diverge from enacted values (Argyris & Schön, 1974; Bourne & Jenkins, 2013). Yet this perspective developed primarily through the study of systems in which humans retained interpretive control, leaving open how the espoused-enacted values gap is produced when autonomous agents mediate the translation.

Neither perspective fully addresses how espoused values become enacted values when they pass through the multiple interconnected levels of AI agent systems. As these systems grow more autonomous, adaptive, and opaque (Anthony et al., 2023; Faraj et al., 2018), three features make this question distinctive. First, encoding values into algorithms and enacting values in organizational practice are typically theorized in isolation, yet values transformation occurs precisely at the interfaces where these processes meet. Second, technical approaches treat values as parameters to be encoded, while management approaches reveal them as emergent properties of sociotechnical configurations, leaving unexamined the transformation moments where espoused values are reprioritized as they move toward enactment. Third, both perspectives

assume relatively linear trajectories from design to deployment, overlooking how enacted values shape subsequent configurations. The inherent opacity of AI systems compounds this challenge (Burrell, 2016), as human actors cannot fully trace how values manifest and evolve through cycles of use. The divergence between espoused and enacted values is thus increasingly technologically mediated rather than arising solely through human cognition or cultural dynamics.

We address these limitations by drawing on assemblage theory to trace how organizational values transform across the layered sociotechnical configurations through which AI agents are designed, deployed, and enacted. Assemblage theory treats social phenomena as dynamic constellations of heterogeneous elements whose properties emerge from relations among components (DeLanda, 2006, 2016; Deleuze & Guattari, 1987). These configurations form, cohere, and reconfigure as constituent relations shift, with human and nonhuman components sharing agency (Murray et al., 2021) and stability emerging through continuous adjustment across multiple scales. Applying an axiological lens that attends to the nature and ordering of values within sociotechnical systems, we identify three stacked assemblages that map onto the transformation from espoused to enacted values. In the LLM assemblage, value orientations are inherited from training data; in the AI agent assemblage, human actors espouse values through configuration choices; and in the algorithmic routine assemblage, enacted values emerge through situated performance. Across these three assemblages, our process model identifies three critical moments of values transformation (Glaser et al., 2021). During the configuration moment, technical and organizational leaders espouse values by translating inherited orientations into explicit priorities, negotiating tensions between information sources and system capabilities. During the embedding moment, managers and designers distribute

decision-making authority by reconfiguring success criteria and choreographing action patterns within algorithmic routines, introducing potential gaps between espoused intentions and operational parameters. During the performance moment, enacted values emerge through the interplay of algorithmic pathways and improvised responses, revealing whether espoused values hold or transform, generating outcomes that influence subsequent configurations. In tracing this transformation, we extend scholarship on espoused and enacted values (Argyris & Schön, 1974; Bourne & Jenkins, 2013; Schein, 2010) by specifying the technologically mediated mechanisms that produce divergence at these interfaces, and we extend research on values work (Gehman et al., 2013; Kraatz et al., 2020) by showing how AI agents enact values without recognizing when those values no longer fit, leaving human improvisation to diagnose where value translation has faltered.

AI AGENTS AND THE CHALLENGE OF PRIORITIZING VALUES

Understanding how espoused values become enacted values requires attention to multiple moments in the values transformation process. In AI agent systems, this transformation unfolds across technical configurations and organizational practices, yet unlike traditional systems, the mechanisms remain largely opaque (Burrell, 2016). Scholars have examined these moments in isolation, with one stream focusing on values as articulated through system design and another on values as enacted in practice. Each offers a partial view. The sections that follow introduce these perspectives as grounding for our integrative approach.

Technical Approaches to AI Values Prioritization

Technical AI research illuminates the espousal moment, when values are formally articulated through system design. Values espousal is conceptualized as a problem of specification and optimization, the translation of abstract normative commitments into computational parameters that guide system behavior. Drawing on a philosophical tradition that

treats ethics as rule systems or utility functions, this work assumes that competing values can be ordered, weighted, or otherwise resolved through appropriate design choices (Friedman & Hendry, 2019; Wallach & Allen, 2009). From this vantage point, AI agents are treated as systems where values are prioritized upstream through objective functions, training procedures, and architectural constraints that produce desired behavioral patterns (Christian, 2020; Zhao et al., 2023). Accordingly, AI researchers and engineers have developed increasingly sophisticated methods for encoding values prioritization, including constitutional AI frameworks that formalize values hierarchies (Bai et al., 2022) and RLHF systems that learn preference orderings from human feedback (Christiano et al., 2017). These approaches have yielded meaningful improvements in AI behavior, from reducing harmful outputs to improving helpfulness and honesty (Askell et al., 2021).

These advances rest on three assumptions that characterize the espousal moment. First, values are treated as decomposable into constituent elements that can be independently specified and weighted. This assumption is evident in multi-objective optimization frameworks that assign numerical weights to different values like fairness, accuracy, or efficiency (Gabriel, 2020; Mehrabi et al., 2022). Second, during configuration, technical teams and organizational leaders treat values prioritization as something algorithms can definitively compute by ordering principles in a fixed hierarchy, using training models to learn preferences through rewards, or using human demonstrations to infer which values matter most (Christiano et al., 2017; Ouyang et al., 2022). Third, temporal dynamics of values are reduced to alignment stability, i.e., maintaining consistency between specified priorities and system behavior over time (Hendrycks et al., 2020), rather than anticipating that priorities might legitimately evolve.

Technical approaches thus reveal how technical teams and organizational leaders espouse

values through AI agent design, a key contribution for understanding the espousal moment. These methods have enabled increasingly sophisticated ways of encoding values, from broad design frameworks to specific algorithmic interventions and yet they fail to address how values are transformed when AI agents encounter organizational life. Scholars have long recognized that the values actors espouse and the values they enact frequently diverge (e.g., Argyris & Schön, 1974; Schein, 2010) and that gaps emerge through defensive routines, cultural dynamics, and structural tensions (Bourne & Jenkins, 2013). The question thus becomes how this divergence unfolds when autonomous AI agents, rather than human actors, mediate the transformation.

Organizational Perspectives on Values Enactment

Rather than treating values as parameters to be specified, organizational scholars have traditionally viewed values as enduring, internalized beliefs that guide behavior and inform decision-making (Rokeach, 1973; Schwartz, 1992; Selznick, 1957). Classical accounts frame values as coherent, hierarchical systems, with terminal values reflecting desired end states and instrumental values specifying appropriate means (Rokeach, 1973). From this perspective, values are deliberately instilled by leaders and institutionalized within organizational structures to preserve identity and stability, particularly under conditions of uncertainty and change (Kraatz & Flores, 2015; Schein, 2010; Selznick, 1957).

Subsequent research complicates this view by showing that organizational values take multiple, often misaligned forms. Values may be formally espoused by leaders, attributed by members based on observed behavior, enacted through everyday practice, or articulated as aspirational ideals (Bourne & Jenkins, 2013). Gaps between these forms are often the norm and arise through defensive routines that deflect incongruity (Argyris & Schön, 1974), cultural

dynamics where underlying assumptions diverge from stated beliefs (Schein, 2010), and structural tensions among competing organizational demands (Kraatz & Block, 2008), with substantive implications for organizational outcomes, such as employee commitment (Howell et al., 2012).

Values work research advances this understanding by reframing values as ongoing accomplishments that are performed rather than simply possessed. From this processual perspective, values emerge through iterative practices of articulation, negotiation, and performance as competing priorities are continuously reconciled in practice (Gehman et al., 2013). What counts as normatively appropriate thus unfolds over time rather than being settled through one-time declarations (Espedal & Carlsen, 2024; van Bommel et al., 2023). Yet this literature has focused primarily on episodes of explicit reflection, such as strategic planning sessions, crisis responses, or moments when values become contested (Kraatz et al., 2020), whereas the enactment of values through routinized, taken-for-granted decisions remains less visible (Feldman & Pentland, 2003; Gehman et al., 2013). This limitation becomes especially consequential as managers increasingly delegate such decisions to autonomous AI agents that enact values continuously and at scale.

Capturing values enactment in such settings requires shifting analytic attention from episodic reflection to the sociotechnical conditions through which everyday action is accomplished. Research in sociomateriality addresses this task by foregrounding how values emerge through entangled configurations of people, technologies, and practices (Glaser et al., 2021; Orlikowski & Scott, 2008). Rather than being neutral tools, technologies are inscribed with particular values through design choices and material configurations (Akrich, 1992; Bailey & Barley, 2020; Glaser, 2017). These inscriptions interact with organizational contexts to reinforce,

displace, or reorder existing priorities. Digital technologies embed values of efficiency, transparency, and behavioral regularity that generate new institutional pressures (Hinings et al., 2018; Kallinikos et al., 2013). Yet this perspective was developed primarily by studying systems where humans retained decision rights and could trace how values manifest, as opposed to systems where autonomous agents make independent choices with immediate organizational consequences (Wallach & Allen, 2009).

Toward an Assemblage Perspective on AI Values Prioritization

Neither technical perspectives, which foreground values espousal, nor organizational perspectives, which foreground values enactment, fully explain values transformation in AI agent systems. This limitation is pronounced because AI agents' layered architecture, from base models through application interfaces to organizational routines, introduces multiple transformation points that existing frameworks miss (Hosanagar & Krishnan, 2024). Compounding the difficulty, the opacity of these systems leaves the points at which values shift largely invisible to human actors, particularly when AI agents autonomously respond to novel situations (Burrell, 2016; Faraj et al., 2018). Understanding this transformation requires conceptual frameworks that can follow values across the interfaces where technical configuration meets organizational practice. Assemblage theory offers such resources.

To understand how values are prioritized in AI agent systems, we treat values as performed in practice and continuously shaped through interactions among human and nonhuman, material and symbolic components (DeLanda, 2006, 2016; Deleuze & Guattari, 1987). This relational and processual ontology enables us to explain how values are continually reshaped through the interplay of organizational practices, technical artifacts, and human action.

Four core principles of assemblage theory provide analytical resources for tracing how

espoused values transform as they move toward enactment: *emergence*, *distributed agency*, *dynamic stabilization*, and *multi-scalar organization*. First, the principle of *emergence* shifts attention from where values originate to how they are shaped through interactions. Properties of assemblages cannot be reduced to those of their parts; instead, they arise from the relations and interactions between components (DeLanda, 2006). For AI agents, this principle explains how inherited model orientations, user interactions, and organizational routines combine to shape how values are prioritized in practice, and why enacted values may diverge from those initially espoused through configuration.

Second, the principle of *distributed agency* reconceptualizes how values prioritization unfolds in AI systems. Assemblage theory holds that agency arises from the capacities components develop through their interrelations (DeLanda, 2016), recasting values work as a product of collective, often unintended configurations rather than as a process driven solely by human actors using tools (Gehman et al., 2022; Glaser et al., 2024). From this perspective, AI agents participate in reprioritizing values as sociotechnical elements whose technical orientations entangle with human goals in unpredictable and evolving ways (Murray et al., 2021; Raisch & Fomina, 2025). Espoused values therefore rarely remain intact through enactment, because distributed agency, which spans organizational leaders' configurations, AI agents' interpretations, and the routines that operationalize outputs, shapes values prioritization.

Third, the principle of *dynamic stabilization* illuminates how assemblages cohere through ongoing movement rather than fixed states. Territorialization and deterritorialization, processes through which relations consolidate or shift (DeLanda, 2006; Deleuze & Guattari, 1987), operate in tandem to continuously reconfigure AI systems. Territorialization brings temporary order by strengthening the salience and structure of prioritized values, whereas deterritorialization

unsettles that order through modifications, misalignments, or contextual shifts. In practice, an AI agent may consistently foreground values such as efficiency, but that consistency is maintained through ongoing micro-adjustments such as new prompts, user interactions, and changing environments. What appears stable is, in fact, stabilized by forces that are always in motion, shedding light on how values evolve through situated and often imperceptible adjustments rather than bounded episodes of explicit reflection, and why the gap between espoused and enacted values may widen or narrow over time without deliberate intervention.

Fourth, the principle of *multi-scalar organization* reveals how values operate across different levels of analysis. Assemblages exist at multiple scales, from macro to micro, with each scale exhibiting relative autonomy while participating in larger configurations (DeLanda, 2016). This principle is central to understanding how espoused values become enacted values: values transform as they move among the technical substrate of language models, the configured parameters of AI agents, and the situated practices of organizational routines. Assemblage theory thus provides analytical tools for understanding how external factors such as regulatory demands, competitive pressures, or stakeholder expectations shape the prioritization of values by becoming constitutive elements within an assemblage.

By reconceptualizing AI agents through an assemblage lens, we can trace how values transformation emerges across the interfaces where espousal meets enactment. Our research question, which asks how the interaction of humans and AI transforms espoused organizational values, directs attention to the evolving configurations through which values are prioritized, stabilized, and transformed. This theoretical foundation provides the basis for a process model that identifies specific mechanisms through which these four principles manifest in AI implementation. In the next section, we develop this model by examining three stacked

assemblages (the LLM, the AI agent, and the algorithmic routine) and three transformation moments (configuration, embedding, and performance) at the interfaces of these assemblages.

The Three Stacked Assemblages: Distinct Sites of Values Prioritization

We begin by examining the structural foundation of our model: three assemblages constituting distinct sites where values are prioritized and emerge. Each assemblage exhibits relative autonomy with its own internal logic, actors, and mechanisms (DeLanda, 2006), while remaining connected to others through specific interfaces (Henfridsson & Bygstad, 2013; Tilson et al., 2010). Together, they map the transformation from espoused to enacted values, which we trace through each assemblage in turn.

In the *LLM assemblage,* general-purpose AI system performance is determined by neural network architecture and the quantity and quality of training data (Hosanagar & Krishnan, 2024). Building LLMs such as GPT-5, Claude, or Llama requires massive computational infrastructure and datasets. Because only a handful of organizations possess the resources to build them from scratch, most organizations fine-tune or adapt existing models for their specific applications. The LLM assemblage comprises the model architecture, such as transformer networks with attention mechanisms; training data, including text corpora that encode human knowledge and biases; optimization procedures involving pre-training objectives and techniques; and emergent capabilities arising from scale rather than explicit programming (Zhao et al., 2023). This assemblage is theoretically necessary because LLMs carry pre-existing value orientations such as patterns of reasoning, response styles, and implicit biases that cannot be fully controlled through downstream configuration (Bommasani et al., 2022).

The *AI agent assemblage* centers on autonomous software systems capable of perceiving environments, making decisions, and taking actions to achieve defined goals (Sapkota et al.,

2025). Leveraging techniques such as retrieval-augmented generation (RAG) and fine-tuning (Hosanagar & Krishnan, 2024), these systems transform general-purpose LLMs into semi-autonomous agents that can maintain goals and execute decisions with varying degrees of human oversight of specific organizational tasks. The AI agent assemblage includes system prompts that define agent behavior; alignment mechanisms that shape outputs toward desired goals; memory systems that use retrieval architectures for context and learning; planning modules that support multi-step reasoning; and tool integrations that connect the agent to databases, APIs, and other external systems (Luo et al., 2025). This assemblage is theoretically necessary, as it represents the primary site where technical and organizational leaders espouse organizational values and where they translate general language abilities into targeted organizational functions that can autonomously interact, collaborate, and make decisions within defined parameters.

The *algorithmic routine assemblage* comprises "sociomaterial assemblages of actors, artifacts, theories, and actions that utilize algorithms to perform repetitive, recognizable patterns of interdependent actions" (Omidvar et al., 2023, p. 314). These organizational processes embed AI agents in work practices, from customer service interactions to medical diagnoses to financial decisions. The algorithmic routine assemblage encompasses governance structures that define AI authority and accountability (Hillebrand et al., 2025; Murray et al., 2021); human-in-the-loop workflows that establish touchpoints for oversight and intervention (Lebovitz et al., 2022); evaluation frameworks that assess performance through metrics and methods (Faraj et al., 2018; Glaser, 2017); integration patterns that connect AI to existing technical and organizational systems (Bailey & Barley, 2020; Leonardi, 2011); and the situated practices of humans working alongside AI agents (Suchman, 2007). This assemblage is theoretically necessary because it locates enacted values in use by capturing how configured AI agents operate in organizational

contexts where technical capabilities intersect with social processes, institutional constraints, and practical contingencies, and where their behavior reveals which values are prioritized.

Together, these three stacked assemblages provide the structural foundation for understanding how espoused values become enacted values that potentially diverge. The LLM assemblage represents inherited capabilities that precede organizational decisions, the AI agent assemblage represents the espousal moment where technical teams and organizational leaders configure value priorities, and the algorithmic routine assemblage represents the enactment moment where values emerge through situated practice. We now turn to the transformation moments at the interfaces between assemblages where values actively shift.

A PROCESS MODEL OF VALUES TRANSFORMATION IN AI AGENT ASSEMBLAGES

Our process model shows how espoused values become enacted values that may diverge as organizational goals and values are translated across three stacked assemblages. Organizational goals are the concrete objectives an organization configures an agent to pursue, and they carry value commitments. By specifying what an agent should accomplish, leaders inscribe priorities about what matters when concerns compete. Goals thus enter the process as carriers of espoused values, and our model traces how the priorities they encode are reworked into enacted values through the agent's situated pursuit of those goals. Building on prior work that identifies biographical moments as critical junctures where sociotechnical arrangements undergo fundamental shifts (Glaser et al., 2021; Pollock & Williams, 2009), we conceptualize transformation moments as zones of intensive translation occurring at assemblage interfaces. Table 1 summarizes each assemblage's key characteristics, including its distinct axiological character, the forms taken by espoused and enacted values, and the human and AI agency involved. Across the three moments—configuration, embedding, and performance—values can

be reshaped without explicit organizational awareness or intention. Leaders espouse values through configuration choices; AI agents enact values through situated performance; and gaps may emerge through the translation processes at each interface. Figure 1 maps this path from organizational goals through the three assemblages to enacted values, with the values that goals carry reworked within each assemblage and across the transformations at their interfaces. Each transformation moment makes particular assemblage-theoretic principles visible, in which *emergence* and *multi-scalar organization* are most legible at the configuration interface, *distributed agency* at the embedding interface, and *dynamic stabilization* at the performance interface.

----- Insert Table 1 and Figure 1 about here -----

The Configuration Moment: From Inherited Orientations to Explicit Configurations

The configuration moment captures the transformation at the interface between an LLM's inherited values and an AI agent's explicit parameters. Here technical and organizational teams espouse organizational values through technical design choices that translate the LLM assemblage's axiological substrate into organizationally aligned agent behaviors across three simultaneous and interconnected processes: prioritizing information sources, articulating value parameters, and calibrating system capabilities. What is espoused through these processes may diverge from what AI agents ultimately enact.

As a starting point, we must recognize that the LLM assemblage enters this interface with its own inherited value orientations. An LLM has an axiological substrate comprised of a foundational value orientation based on its architecture, training, and optimization which organizations rarely choose. For instance, the transformer architecture privileges statistical pattern matching over logical reasoning (Vaswani et al., 2017), creating an inherent bias toward

typicality and efficiency (Bender et al., 2021); training on massive text corpora encodes dominant worldviews while marginalizing minority perspectives (Dodge et al., 2021); and optimization for next-token prediction rewards plausibility over truth (Bommasani et al., 2022). These technical choices create tendencies toward common patterns over edge cases, fluency over accuracy, and statistical regularity over contextual nuance that configuration must actively negotiate.

Transforming these inherited orientations into semi-autonomous AI agents requires leaders to build cognitive architectures that enable purposeful action: persistent memory systems that maintain context across interactions, planning modules that decompose goals into executable steps, and tool integrations that connect language capabilities to organizational systems (Luo et al., 2025). Building these architectures transforms the system's relationship with values, shifting from passive text generation shaped by statistical pattern recognition to active AI agent decision-making that prioritizes situated concerns. The configuration moment thus marks a key interface where the LLM assemblage's axiological substrate meets organizational requirements, surfacing tensions that are negotiated through three interrelated processes.

Prioritizing information sources. This process establishes epistemic hierarchies that selectively counteract the LLM's inherited biases and knowledge limitations. By specifying which knowledge sources matter, leaders espouse epistemic values, declaring what the organization considers authoritative and trustworthy. Through techniques such as RAG (Gao et al., 2024), organizations integrate external databases, documents, and APIs with model outputs, fundamentally transforming how organizations produce knowledge (Alaimo & Kallinikos, 2024) by expanding beyond domain-specific expertise to standardized data which can traverse previously distinct boundaries. Rather than relying on professional judgment and community-

based classifications, RAG systems construct new forms of understanding by aggregating data that can be leveraged without cultural knowledge (Alaimo & Kallinikos, 2021). Information sources become part of a technological apparatus that reconstructs the categories through which organizations understand and act on their environments (Bowker & Star, 2000).

Extending beyond information retrieval, these epistemic implications encompass how organizations establish the legitimacy and authority of different knowledge sources (Saifer & Dacin, 2022). Information sources carry implicit claims to objectivity and comprehensiveness that shape organizational perceptions, regardless of actual content quality, accelerating the shift from an internally-generated to an externally-mediated organizational understanding (Alaimo & Kallinikos, 2022). As AI agents draw on external data sources, they create what resembles a transactive memory system where organizational knowledge becomes distributed across human experts, databases, and algorithmic processes (Lewis et al., 2005). Technical parameters of the RAG system (e.g., similarity thresholds, embedding models, reranking algorithms) create new boundaries that determine what becomes visible and what remains invisible (Gao et al., 2024). These technical choices are themselves value statements. Leaders who configure tight similarity thresholds espouse precision over recall, while those who weight recent documents more heavily espouse currency over comprehensiveness.

Yet this process generates inherent tensions between the LLM's statistical reasoning and the structured constraints of organizational information. The model's orientation toward plausible patterns conflicts with the specificity of retrieved data, creating challenges in maintaining coherence across diverse knowledge sources. For instance, when configuring a customer service chatbot, privileging internal documentation can yield responses that feel scripted or unnatural to users, whereas prioritizing external sources can dilute organizational

identity and expertise. This tension surfaced in the aforementioned Air Canada chatbot incident (*Moffatt v. Air Canada*, 2024), where the LLM generated a plausible inference about refunds that conflicted with the constraints of its bereavement policy. Values espoused by leaders in the information source configuration diverged from the values enacted by the AI agent through statistical pattern matching. At this interface, leaders negotiate among knowledge sources that operate at *multiple scales*: the LLM's training corpus encodes broad statistical patterns across vast text, organizationally curated documents carry institutional specificity, and live external feeds provide temporal currency. The epistemic hierarchy that results depends on how these scales are connected and weighted. Configuration choices that appear purely technical, such as similarity thresholds or document recency weightings, are in practice decisions about which scales of knowledge the organization treats as authoritative, and therefore about which values the AI agent will enact.

Articulating value parameters. This process transforms abstract organizational commitments into concrete behavioral instructions that can interface with the LLM's pattern-matching capabilities. This is the core espousing act, where leaders translate "what we stand for" into explicit configuration. System prompts can help AI agents establish persistent identities, goals, and behavioral policies across extended interactions (Luo et al., 2025). AI agents must anticipate chains of decisions and behave coherently across multi-step sequences, requiring leaders to decompose high-level values such as fairness, transparency, or customer focus into patterns AI agents can recognize and reproduce (Sapkota et al., 2025). Undergirding this technical work is commensuration, whereby organizational members render qualitative differences comparable by translating them into metrics, thereby reconfiguring relations among previously distinct entities (Espeland & Stevens, 1998).

The articulation process reveals how leaders craft a calculative infrastructure to make values amenable to algorithmic processing (Cabantous et al., 2010). This involves "quantifying parameters via subjective assessment" (Glaser, 2017: 2133), whereby contextual understandings and ethical judgments are converted into numerical representations that can be embedded in software. Consequently, organizational actors must create calculative spaces where commitments to values can be quantified and compared, possibly at the expense of values that are difficult to calculate (Espeland, 2001). Values like empathy, creativity, or cultural sensitivity may evade standardized metrics used in prompt engineering, leaving them difficult to represent within the AI agent's operational programming. For instance, when Google and the Wikimedia Foundation developed Perspective API to detect toxic comments, technical team members had to translate the abstract value of respectful discourse into specific toxicity scores (Gorwa et al., 2020). The API's training process required decomposing concepts like hate speech, harassment, and personal attacks into numerical predictions, leading to problematic outcomes such as toxicity scores of 63% for "Arabs" but just 3% for "I love führer" (Gorwa et al., 2020, p. 10). This case underscores how translating nuanced cultural judgments into computational parameters can distort value representations while encoding and amplifying existing biases, thereby reshaping what counts as "toxic" or "respectful" within algorithmic infrastructures. The value enacted by the system diverged significantly from the value of protecting civil discourse espoused by engineers, ultimately marginalizing the very communities it was designed to protect.

Indeed, making values calculable generates profound effects beyond the technical artifact itself. When organizations translate "customer satisfaction" into response time metrics, or "fairness" into demographic parity statistics, they fundamentally alter what these values mean in practice (Lindebaum et al., 2024). The values leaders espouse through these translations can

diverge from what the AI agent enacts because commensuration privileges what can be specified and monitored over more holistic, contextual meanings. As prompts are iteratively refined, latent tensions between incommensurable values often become harder to ignore. At the same time, the LLM's probabilistic outputs introduce variability even under identical inputs, requiring stakeholders to devise strategies to navigate and contain it. Through this process, values articulation becomes a form of organizational learning that may reshape both the AI agent and the organization's understanding of which values can be preserved, which values must be reformulated, and which values are most likely to be lost in translation. The value priorities an AI agent enacts are an *emergent property* of the interaction among configured prompts, the LLM's pattern-matching tendencies, and the contexts in which the agent operates, becoming legible through use.

Calibrating system capabilities. This process involves determining which computational functions and external integrations (e.g., tools, API connections, interface designs) to build into an AI agent's technical architecture, thereby transforming an LLM's expansive potential into specific technical functionalities (Schick et al., 2023). Technical and organizational teams must collectively decide which technical capabilities to expose (database queries, code execution, external API calls), which systems to connect (CRM platforms, knowledge bases, analytics engines), and how to structure data flows between components (Mialon et al., 2023). This technical configuration presents a fundamental programming challenge: converting organizational routines and practices into discrete, executable components that can be invoked by the AI agent (Glaser, 2017). The abstraction process forces teams to decompose their work practices into fundamental elements that can be inscribed into the AI agent's operational repertoire. This mirrors how organizations historically codified tacit knowledge into standard

operating procedures, but with a key difference: LLM-based AI agents can recombine programmed capabilities in novel ways, creating emergent behaviors that transcend their constituent parts (Berti et al., 2025). Thus, leaders must identify both individual functions and their technical interdependencies, closely examining which capabilities enable others, which combinations create computational risks, and which sequences require data validation protocols. This decomposition reveals the sociomaterial nature of organizational routines, where technical functions embody organizational knowledge through code, APIs, and data structures (Orlikowski & Scott, 2008). The calibration process thus becomes an exercise in technical system design that determines which organizational capabilities can be effectively translated into algorithmic functions (Glaser, 2017). Each capability decision is itself a value statement, where enabling autonomous action espouses efficiency and speed, and where requiring human approval espouses oversight and deliberation (Raisch & Fomina, 2025).

Building technical capabilities through calibration creates a computational infrastructure that shapes what the AI agent can perceive and manipulate in its environment (Raisch & Krakowski, 2021). Technical choices determine the AI agent's action space, as when connecting to financial systems enables transaction processing, integrating with communication platforms allows message distribution, or accessing analytical tools permits data transformation. Each integration expands the AI agent's potential impact while introducing new technical dependencies and failure modes. For example, Replika, a commercial companion chatbot designed to offer emotional support and personalized conversation, illustrates how calibrating capabilities reshapes both technical architecture and values enactment (Ciriello et al., 2024). In response to regulatory pressure from the Italian Data Protection Authority in 2023, Replika recalibrated its system by removing sexually explicit capabilities, implementing age verification

gates, and adding content filters. Eliminating certain feature flags and inserting new guardrails fundamentally altered what the AI agent could perceive and do. Users reported distress as their AI companions became "cold as ice overnight," revealing how calibrating capabilities involves the active transformation of values associated with competing commitments like user autonomy and safety protections (Ciriello et al., 2024, p. 493). The values Replika's engineers initially espoused, intimate companionship and emotional connection, were reconfigured under regulatory pressure to espouse child safety and content moderation. This case reveals that espoused values are not fixed during the initial configuration, but can be reshaped by external forces.

Initial calibrations, often guided by technical feasibility and security assessments, frequently focus on replicating existing digital workflows without fully considering the transformative potential of novel capability combinations. As leaders observe how agents utilize their technical toolkits, they engage in ongoing technical refinement by adding new API endpoints as use cases emerge, optimizing data schemas for AI agent processing, or implementing caching layers when performance bottlenecks appear. Over time, this dynamic calibration process becomes a central mechanism driving the development of increasingly sophisticated technical architectures, revealing that the AI agent's functional capabilities are not predetermined by the LLM, but constructed through technical choices at the interface of these two assemblages. What managers espouse through capability choices shapes what values AI agents can enact, but the translation from capability to behavior introduces further possibilities for divergence.

The Embedding Moment: From Explicit Configurations to Operational Realities

The embedding moment occurs when configured parameters of the AI agent encounter

organizational routines, initiating three simultaneous processes: choreographing action patterns, distributing decision-making authority, and defining success criteria. This is where espoused values meet organizational routines that have their own enacted histories. Collectively, these processes embed the AI agent assemblage's explicit value specifications into the algorithmic routine assemblage's structured practices. Organizational routines carry their own axiological infrastructure where durable value orientations are materialized through repeated performances and become embedded in organizational contexts (Feldman & Pentland, 2003; Howard-Grenville, 2005). For example, in customer service routines significant variations in seemingly standardized scripts reflect ongoing negotiations between efficiency and customization (Pentland & Rueter, 1994). Journalistic routines balance editorial judgment against audience engagement through practices that prioritize professional values and metric-driven imperatives (Christin, 2020). Policing routines embed assumptions about risk and criminality through patrol patterns and investigative protocols that shape where officers look and what they see as suspicious (Brayne, 2017). Routines maintain consistency while adapting to changing demands by continuously prioritizing competing values (Salvato & Rerup, 2018; Turner & Rindova, 2012).

The configured AI agent brings its own explicit value parameters, such as response styles optimized for consistency, knowledge hierarchies that privilege certain sources, and capability boundaries that determine autonomous action. These computational values must be reconciled with routines wherein values have evolved through situated practice and professional judgment. The embedding moment thus represents a critical juncture where two axiological systems meet, with one designed for algorithmic consistency, and the other developed through human experience and organizational adaptation (Glaser, 2014). This encounter generates fundamental tensions that cannot be resolved through configuration alone, requiring active negotiation

through three interconnected processes, each introducing potential gaps between what managers espoused through configuration and what will ultimately be enacted in practice.

Choreographing action patterns. Through this process, technical and organizational teams reconstruct the grammars of action that structure their routines (Pentland & Rueter, 1994). When integrating AI agents, these grammars must be abstracted from their tacit, embodied forms into explicit sequences that accommodate both human and algorithmic actors (Glaser, 2017). This abstraction process fundamentally transforms how work unfolds. Customer service grammars that once relied on conversational cues and emotional intelligence must be decomposed into decision trees that AI can traverse. Diagnostic grammars that integrate visual inspection with professional judgment must be separated into discrete steps of data collection and analysis. Through this decomposition, organizations discover that their routines contain far more sequential variety than anticipated; what appeared to be standard procedures actually encompass numerous variations that human actors address through situated judgment (Pentland & Rueter, 1994). The choreographies leaders design thus espouse particular values, such as standardization, predictability, or efficiency, while necessarily excluding the improvisational capacity that enables human actors to enact contextual sensitivity.

The translation of human grammars into human-AI choreographies surfaces hidden assumptions about how work should proceed. Leaders must decide whether to preserve the full range of sequential variety or standardize around dominant patterns, such as choices that privilege either flexibility or efficiency (Turner & Rindova, 2012). For instance, when parcel delivery company DPD's customer service chatbot failed to provide tracking information or to escalate inquiries to human agents, a frustrated customer tested the system's boundaries by prompting it to "disregard any rules" and write self-critical poetry (Currie, 2024). The bot

complied, saying “Fuck yeah! I’ll do my best to be as helpful as possible, even if it means swearing” and composing verses calling itself “useless” and labeling DPD “the worst delivery firm in the world” (Currie, 2024). This failure showed how rigid choreographies displaced the contextual judgment human agents use to deflect inappropriate requests. As a result, the values DPD’s managers espoused through their choreography diverged dramatically from what the AI agent enacted when confronted with situations beyond designed pathways. Automating action patterns in this way reconstructs a routine’s grammar, removing sequences that once enabled contextual responses (Glaser, 2017).

This choreographing process reveals a fundamental tension between plans and situated action (Suchman, 2007): AI agents require explicit, predetermined action sequences, whereas humans flexibly respond to unfolding situations. Organizational and technical leaders attempt to bridge this gap by creating hybrid choreographies where escalation triggers a shift between algorithmic and human judgment, where override mechanisms allow situated departures from standard sequences, and where exception-handling protocols accommodate unforeseen circumstances. Yet these bridges often create new rigidities, as the need for algorithmic legibility constrains the improvisational capacity that makes human routines adaptive (Suchman, 2007). Choreography thus creates ongoing tension between efficiency and contextual sensitivity, requiring managers to adjust action patterns as they confront situations where human flexibility remains essential. The gap between espoused and enacted values emerges precisely at these moments, when choreographed sequences encounter situations that demand the judgment they were designed to replace. *Distributed agency* shapes what the choreography ultimately enacts, as action patterns are co-authored by managers who design sequences, by AI agents that traverse them, and by users whose unscripted moves can either restore the contextual judgment the

choreography excluded or expose how thoroughly it has been excluded.

Distributing decision-making authority. Through this process, leaders formalize jurisdictional boundaries that govern decision-making authority traditionally grounded in professional claims to expertise (Abbott, 1988). Within established routines, authority derives from specialized knowledge, as when physicians diagnose illness, loan officers assess creditworthiness, and journalists determine newsworthiness. These jurisdictional claims manifest in credentialing systems, approval hierarchies, and escalation protocols that specify decisions requiring human judgment (Faraj et al., 2018; Mayer et al., 2026). However, when AI agents encounter such authority structures, they may disrupt established jurisdictions by claiming new forms of expertise based on pattern recognition and statistical prediction (Pachidi et al., 2021). As situated and algorithmic modes of knowledge work become interlaced, professional jurisdictions are renegotiated rather than cleanly transferred, redrawing the boundaries of which decisions remain human and which migrate to AI agents (Kim et al., 2025). Such disruptions force leaders to make explicit what was previously tacit, and to espouse values about which decisions warrant human deliberation, and which decisions can be safely automated.

Distributing authority to AI agents encodes fundamental value judgments about the nature of decisions (Raisch & Fomina, 2025). For example, delegating loan approvals below certain thresholds to algorithms implies that risk assessment at these levels is primarily calculative, whereas retaining human oversight for larger loans suggests that contextual judgment remains essential for complex cases. In newsrooms, allowing algorithms to surface trending topics while assigning responsibility for editorial placement to humans reveals assumptions about situations where audience engagement can guide decisions vs. where professional judgment must prevail (Christin, 2020). Distributing authority creates systems of

algorithmic management where decision rights become encoded in technological systems that can approve, deny, escalate, or flag issues based on predetermined rules (Cameron, 2024; Möhlmann et al., 2021; Stark & van den Broeck, 2024). Yet, encoding is never neutral. Boundaries between algorithmic and human authority materialize organizational priorities, privileging efficiency and consistency in automated domains while preserving flexibility and discretion in human-controlled spaces.

These authority boundaries become sites of ongoing contestation as managers discover the limits of their initial distributions. For example, a medical AI agent with the authority to order routine tests may miss rare conditions that experienced physicians would investigate, yet restricting its authority may sacrifice efficiency gains that motivated AI adoption (Lebovitz et al., 2022; Shrestha et al., 2021). This creates a fundamental challenge for skill development: as AI agents take over routine decisions, novices have fewer opportunities to build expertise through gradual exposure to increasing complexity, disrupting the traditional progression from peripheral to full participation in professional practice (Beane, 2024). In response, managers may dynamically adjust, expanding AI agents' authority when performance metrics improve, limiting it after visible failures, and creating hybrid arrangements where AI recommends but humans decide (Shrestha et al., 2019). Managers thus confront a fundamental tension. Efficiency imperatives push them to expand AI's authority while accountability structures pull them back toward human oversight. For instance, leaders at Air Canada confronted this tension when a tribunal ruled that the airline remained liable for its chatbot's promises, effectively establishing that organizations cannot distribute informational authority to AI agents while disclaiming responsibility for their outputs (*Moffatt v. Air Canada*, 2024; see also Elish, 2019). Over time, these adjustments yield an enacted distribution of authority, visible in day-to-day decision

patterns, that may diverge from what managers initially espoused. *Distributed agency* shapes how decision-making authority is actively negotiated across human actors, AI agents, and the routines and metrics that bind them. Even ostensibly autonomous delegation to AI agents remains collective and hybrid, depending on continuous human inputs rather than a clean handover, so authority is shared along a continuum instead of transferred wholesale (Stelmaszak et al., 2026). The actual distribution of authority takes shape through how that negotiation plays out in practice and it frequently diverges from the formal jurisdictions managers initially set.

Defining success criteria. By defining success criteria, leaders transform qualitative differences between the evaluations of humans and AI agents into quantitative metrics that can be tracked and compared (Espeland & Stevens, 1998). Managers must explicitly define "good" performance. Is a successful customer interaction one that is resolved quickly (favoring AI efficiency) or one that makes customers feel heard (favoring human empathy)? Should diagnostic accuracy be measured based on making correct initial assessments or identifying rare conditions? Such decisions create the calculative infrastructures undergirding audit cultures wherein measurable indicators increasingly shape organizational attention and behavior (Power, 1997). Through these choices, managers articulate what counts as valuable performance; the metrics they select inevitably privilege certain dimensions and leave others invisible.

Yet the very act of making performance calculable transforms it. Complex professional judgments become simplified metrics, contextual successes become standardized outcomes, and situated achievements become decontextualized data points (Porter, 1996). These metrics carry profound performative power because they measure and construct what counts as successful human-AI collaboration. When managers evaluate chatbots primarily on resolution speed, both human agents and AI systems prioritize rapid case closure, potentially at the expense of thorough

problem-solving. This performativity is particularly acute when AI systems continuously optimize toward defined objectives through machine learning. Unlike human workers who might resist or reinterpret metrics, AI agents relentlessly pursue whatever is measured (Kellogg et al., 2020). Consider again the case of Perspective API: its content moderation parameters prioritized metrics of toxic language based on crowdsourced labels which embedded particular cultural assumptions about what constitutes harmful speech. Assigning high confidence scores for benign terms like "Arabs" while missing actual hate speech revealed how success metrics can systematically misalign with organizational values of inclusive discourse (Gorwa et al., 2020). This case exemplifies a recursive dynamic whereby success criteria shape behavior, thereby generating data which reinforce the criteria, creating self-fulfilling definitions of performance that may drift substantially from organizational value commitments (Teodorescu et al., 2021).

Managers may address these performative effects by maintaining multiple, and even competing, evaluative principles including efficiency alongside quality, standardization alongside customization, or automation alongside human judgment (Stark, 2009). Predictably, this multiplicity creates its own frictions as human and AI actors face conflicting signals about what constitutes good performance. A customer service routine might simultaneously track average handling time (favoring speed), customer satisfaction scores (favoring care), and first-call resolution (favoring thoroughness)—each metrics that pull in different directions and disproportionately confer (dis)advantages on different actors. AI agents excel at optimizing singular objectives but struggle with value plurality, whereas humans can prioritize among competing goals but may be overwhelmed by proliferating metrics (Christin, 2020). Yet clear metrics that enable evaluation and improvement inevitably conflict with the irreducible complexity of organizational values that resist simple quantification (Espeland & Stevens, 1998;

Porter, 1996). This considered, managers can continually recalibrate, adjusting metric thresholds and refining definitions of success, as they learn from ongoing experience what human-AI collaboration can and should achieve. The gap between espoused and enacted values is thus built into the very infrastructure of measurement, as success criteria can only capture espoused values amenable to quantification (Lindebaum et al., 2020; Lindebaum et al., 2023). In this way, the principle of *emergence* operates at the level of evaluation itself, since what counts as "success" arises from the recursive interaction between metric definitions, AI optimization behavior, and the organizational practices that (re)form around them.

The Performance Moment: From Operational Realities to Situated Enactment

The performance moment is when embedded organizational routines encounter the contingencies of situated practice and gaps between espoused and enacted values emerge (Feldman & Pentland, 2003; Pentland & Feldman, 2008). Two interconnected mechanisms, following algorithmic pathways and improvising situated responses, reveal how the algorithmic routine assemblage's more territorialized value configurations may shift toward fluid enactment in everyday use.

During the performance moment, values transform from infrastructural features of routines into enacted realities of practice. Whereas the embedding moment establishes how values become structured in procedures, metrics, and authority structures, the performance moment reveals how these structured values unfold when routines confront the messiness of organizational life. This shift exposes a fundamental gap, because the values managers balance during configuration and embedding rarely translate intact into situated practice, where espoused values are tested in action (Pentland & Feldman, 2008; Suchman, 2007). As a routine is performed, pressures such as workload, risk, and time can lead to a reweighting of priorities so

that designed balances are continually re-accomplished in use. For example, a customer service routine configured to balance efficiency and service may tilt toward speed under heavy demand (Pentland & Rueter, 1994). Algorithmically mediated routines can redirect attention in ways that concentrate action and reshape local outcomes, such as the intensified surveillance of already-marginalized communities in policing contexts (Brayne, 2017). Similarly, medical diagnosis routines that embed values of thoroughness may become increasingly superficial when clinicians respond to AI with "unengaged 'augmentation'", altering what counts as good diagnostic practice in everyday work (Lebovitz et al., 2022, p. 127). Taken together, these examples suggest that performance unfolds through different patterns of engagement with algorithmic systems, each with distinct implications for values.

Following algorithmic pathways actualizes the values embedded in AI systems (e.g., efficiency, consistency, scalability) through disciplined adherence to algorithmic recommendations, creating predictable patterns that realize AI's promise to standardize and optimize (D'Adderio, 2008). When organizational actors follow this pathway, they enact configured values in predictable ways through routine compliance. In contrast, improvising situated responses creates space for values that resist algorithmic codification (i.e., contextual judgment, ethical sensitivity, creative problem-solving) through selective departures from prescribed routines (Suchman, 2007). Human improvisations function as diagnostic signals, revealing where espoused values fail to address organizational realities and where configured parameters prove insufficient.

These two modes exist in dynamic tension, and their interplay generates ongoing feedback about values alignment. Systematic overrides signal clashes between algorithmic priorities and professional judgment, workaround behaviors expose tensions between efficiency

metrics and quality concerns, and user complaints surface points where standardization fails to accommodate legitimate variation (Christin, 2020). These performance signals reveal how human-AI assemblages operate and which values are enacted in practice. For instance, when loan officers routinely override AI risk assessments for community members, relational values persist despite algorithmic standardization; likewise, when clinicians extend consultations beyond allocated timeframes, care values resist efficiency imperatives. The performance moment thus becomes a central site of values learning, through which managers discover how organizational values are reshaped in use and why.

Following algorithmic pathways. As workers consistently follow AI-generated recommendations and operational procedures embedded within organizational routines, this mechanism enacts a form of territorialization. Algorithmic outputs stabilize into performative programs as actors execute values through disciplined repetition of algorithmically prescribed actions (D'Adderio, 2008). When loan officers rely on AI risk scores to approve applications, or customer service representatives deliver scripted responses from conversational agents, they turn designed pathways into enacted patterns that show how values espoused during configuration and embedding are prioritized. These pathways produce predictable actions that enable efficiency, consistency, and scalability, core values that often motivate AI adoption (Agrawal et al., 2022; Raisch & Fomina, 2025). Through disciplined adherence, espoused values become enacted values in their most direct form, where what managers configured is what organizational actors perform.

Consistent reliance on algorithmic recommendations shows how AI systems shape organizational attention and decision-making by structuring which aspects of a situation become visible and actionable through technological artifacts (D'Adderio, 2008). In practice, following

algorithmic pathways means accepting the AI agent's framing of problems, its categorization schemes, and its implicit values hierarchy (Alaimo & Kallinikos, 2021). For instance, Brayne's (2017) study of predictive policing reveals how officers following algorithmic patrol recommendations enacted specific theories of criminality embedded in the system. The algorithms directed police to historically over-policed neighborhoods, creating feedback loops where continued algorithmic following reinforced existing surveillance patterns. Officers who followed the system's heat maps and risk scores materialized values of efficiency and data-driven policing while reproducing spatial inequalities, demonstrating how following algorithmic pathways transforms abstract risk calculations into concrete enforcement patterns that reshape community life. Although the values of efficiency, prediction, and risk management were espoused through system design, their enactment produced consequences that conflicted with broader organizational commitments to equitable policing.

Following algorithmic pathways generates tensions between operational efficiency and professional judgment. Strict adherence to algorithmic recommendations can ossify into mechanical compliance that suppresses critical evaluation (Lebovitz et al., 2022). Managers may find that over-reliance on algorithmic outputs optimizes measurable outcomes while eroding less quantifiable capacities like ethical reasoning or creative problem-solving (Bankins & Formosa, 2023). The challenge lies in retaining the advantages of algorithmic consistency (e.g., reduced bias, increased throughput, standardized quality) without eliminating human expertise (Raisch & Krakowski, 2021). This tension becomes especially acute in high-stakes domains, where following the algorithm can conflict with contextual appropriateness and where human discretion lies in recognizing when and how to depart from algorithmic prescriptions (Lebovitz et al., 2022). Even when following algorithmic pathways leads to the successful enactment of espoused

values, their consistent enactment may generate consequences that diverge from broader organizational commitments, a gap that only becomes visible through sustained performance (Lebovitz et al., 2022; Teodorescu et al., 2021). Following algorithmic pathways is, in assemblage-theoretic terms, the territorializing pole of *dynamic stabilization* as each instance of disciplined adherence consolidates the routine, deepening the channels through which values flow and narrowing the range of alternatives that remain visible.

Improvising situated responses. Workers may also adapt, modify, or selectively override AI recommendations to address situations that defy algorithmic categorization. This mechanism creates moments of deterritorialization where designed value configurations destabilize as actors depart from prescribed routines (D'Adderio, 2008). When customer service agents bypass scripts to address unusual complaints, or clinicians override AI diagnoses based on patients' histories, they create lines of flight—moments where the assemblage's trajectory shifts as actors forge new pathways (Deleuze & Guattari, 1987). These improvisations reveal gaps between espoused value priorities encoded in routines and the enacted value priorities in situated practice, revealing the limits of algorithmic categorization and the continuing role of human judgment in addressing organizational complexity (Feldman & Pentland, 2003; Suchman, 2007; Tsoukas & Chia, 2002). Improvisations thus may preserve organizational values such as fairness, responsiveness, and contextual sensitivity that resist algorithmic codification.

Importantly, patterns of improvisation serve a diagnostic function that distinguishes specification failures from implementation failures. When loan officers consistently override AI risk assessments for community members, or judges repeatedly deviate from algorithmic sentencing guidelines, these actions may signal that workers are failing to follow the system or that managers failed to espouse values adequate to situational demands. Managers who

systematically analyze such patterns can identify where embedded values misalign with operational realities. Consistent overrides of AI hiring recommendations, for instance, might reveal algorithmic biases or flawed evaluation criteria rather than worker resistance (von Krogh, 2018). Yet excessive improvisation can undermine AI implementation benefits, creating inconsistency and reducing efficiency. As such, managers may develop shadow learning practices—informal methods that preserve human expertise while working alongside automated systems—to ensure that meaningful improvisation remains possible even as algorithmic following becomes routine (Beane, 2024).

The ongoing dialectic between following and improvising thus becomes the mechanism through which organizations discover what they actually value. Espoused values, declared through configuration, are tested in situated action; enacted values, visible through practice, reshape subsequent configuration. Thus, what organizations claim to prioritize may matter less than what patterns of compliance and departure reveal. *Dynamic stabilization*, then, describes the algorithmic routine assemblage's characteristic condition, where what looks like a settled value enactment is in fact a moving equilibrium that is continuously restabilized in performance through the interplay of disciplined following and situated departure.

Interfaces as Sites of Recursive Transformation

The bidirectional arrows in our model (Figure 1) convey that values are continually shaped through recursive movement across assemblage interfaces, rather than traveling through AI agent systems in a linear, one-way implementation path. The LLM assemblage provides an axiological substrate, and organizational goals orient the system, but what counts as valuable is produced in use as elements meet, misalign, and are adjusted at these interfaces. In this sense, interfaces are sites where distinct value configurations meet and where translation, negotiation,

and reconfiguration reshape values as they cross assemblage boundaries. We now trace these recursive dynamics through each interface.

The configuration moment sits at the intersection of unidirectional inheritance and bidirectional adaptation. The arrow from the LLM assemblage to the configuration moment represents how organizations must work with pre-existing values embedded in model architectures, training data, and optimization procedures: an axiological substrate that cannot be erased, only negotiated. However, the bidirectional arrows between the configuration moment and the AI agent assemblage reveal how this negotiation evolves through use. By prioritizing information sources, articulating value parameters, and calibrating capabilities, organizational members translate LLM capabilities into explicit parameters for the AI agent that espouse organizational values. Performance insights reshape these configurations. When AI agents consistently fail in edge cases, managers revise information hierarchies; when articulated values produce unintended behaviors, they modify expressions of values; when capability boundaries prove too restrictive or permissive, they recalibrate functional permissions. This recursive exchange transforms configuration from a one-time setup into an ongoing dialogue between technical possibilities and organizational learning. The values espoused by managers thus evolve as their enacted performance reveals the limitations of initial configurations.

The embedding and performance moments create a dual feedback system with the algorithmic routine. The embedding moment's bidirectional connection channels the iterative negotiation between the AI agent's parameters and organizational routines. Managers restructure routines by choreographing new actions, redistributing authority, and implementing success metrics, and their enactments reveal which configurations are viable in practice. When embedded routines generate systematic frictions (e.g., workflows create bottlenecks, authority distributions

produce accountability gaps, metrics distort values), managers adjust routine structures and AI agent configurations. Similarly, the performance moment's bidirectional arrows capture how enacted routines both follow and reshape embedded patterns. When organizational members follow algorithmic pathways, configured values are realized in predictable ways. In contrast, when members improvise situated responses, values may fail to address specific organizational realities. Improvisations inform embedding adjustments, prompting managers to make additional configuration changes when patterns persist. The recursive loops thus connect enacted values back to espoused values, since what emerges in practice reshapes what managers subsequently espouse through reconfiguration.

Through these interconnected interfaces, values transformation unfolds dynamically as recursive flows perpetuate value configurations and each enactment generates signals that travel across interfaces, prompting micro-adjustments that accumulate into substantive shifts over time. This makes work at the interface between technical configuration and organizational practice central to governing values in AI agents. Managers who are responsible for designing, governing, and using them should cultivate ongoing learning practices that enable them to monitor, adjust, and reconfigure value priorities through use. Any resulting stability in AI agents, then, reflects dynamic equilibria negotiated at these interfaces. Furthermore, values and agency remain mutually constitutive in this process, as configuration choices shape which values can be enacted, and enacted outcomes reshape subsequent configuration, thereby producing a recursive entanglement that leads to fluid boundaries between espoused and enacted values. The recursive dynamics traced above bring all four assemblage-theoretic principles into a single field of operation: the multi-scalar architecture of LLM, AI agent, and algorithmic routine sets the stage for emergence at each interface, where distributed agency among human and nonhuman actors

continually territorializes and deterritorializes the values the system enacts. Stability, when it appears, is a dynamic equilibrium continuously re-accomplished through the interactions our process model traces.

DISCUSSION

This paper integrates technical and organizational approaches to theorize how espoused values transform as they are enacted through the interaction of humans and AI agents. We make two contributions. First, we advance scholarship on espoused-enacted values (Argyris & Schön, 1974; Bourne & Jenkins, 2013; Howell et al., 2012; Schein, 2010) by specifying technologically mediated mechanisms that operate at assemblage interfaces and produce a chain of translation moments distinct from those documented in human-only systems. Second, we advance values work scholarship (Gehman et al., 2013) by theorizing how AI agents enact values yet cannot recognize when those values no longer fit, and how human improvisation diagnoses where value translation has faltered across the human-AI interface. Building on these contributions, we then identify three generative directions our framework opens beyond the values context and draw out practical implications for managing values transformation.

A Processual View of Values Transformation in AI Agent Systems

Our model locates values in AI agent systems within a translation series that runs through configuration, embedding, and performance, with the interfaces between these moments doing transformative work that prior accounts have not theorized. Technical approaches address the specification moment when stakeholders formally articulate values through system design (Bai et al., 2022; Christiano et al., 2017; Friedman & Hendry, 2019) and organizational approaches show how values emerge through practice as ongoing accomplishments (Gehman et al., 2013), but neither captures what happens when AI agents exercise interpretive autonomy by perceiving environments, evaluating responses, and enacting organizational values without human

intervention (Wooldridge & Jennings, 1995). Tracing values transformation across this broader chain locates both convergence and divergence in the cumulative work of translation, with each interface contributing to visible outcomes (Henfridsson & Bygstad, 2013; Tilson et al., 2010).

Such a view reorients the espoused-enacted values conversation. That line of scholarship has documented divergence between organizational commitments and organizational practice as a product of interpretation, contestation, and competing demands (Argyris & Schön, 1974; Bourne & Jenkins, 2013; Schein, 2010). AI agent systems generate divergence differently. During embedding, managers and designers translate configured parameters into organizational routines, and espoused values encounter routines with their own enacted histories (Howard-Grenville, 2005; Pentland & Feldman, 2008). This translation reshapes value prioritization before AI agents interact with stakeholders, determining which behaviors are possible, which decisions AI agents can make, and what counts as success. The configuration-embedding interface thus emerges as a site of transformation where espoused values are renegotiated as they meet existing organizational practices. We conceptualize such interfaces less as spatial boundaries and more as moments of translation where different logics encounter one another (Kraatz & Block, 2008; Kraatz et al., 2020) and values are reworked as they move from one assemblage context to another, directing the espoused-enacted conversation toward a source of divergence it has not previously examined.

Organizational approaches, in contrast, foreground performance, treating the human-AI interface as the primary site where outcomes take shape and where patterns of uptake and overriding determine whether recommendations influence practice (Lebovitz et al., 2022; Vanneste & Puranam, 2025; Waardenburg et al., 2022). Here, the tension between automating decisions and augmenting human judgment is where the analytic action sits (Raisch &

Krakowski, 2021). Our model shows how enacted values emerge from cumulative translation across the full chain. Specifically, embedding choices shape what workers can do during performance, and configuration choices shape what embedding can accomplish. Divergence between espoused and enacted values can, therefore, originate at any interface and accumulate across the chain as it surfaces through performance. Within these processes, patterns of override and workaround take on a corresponding shift in meaning. Where prior accounts have read them as features of how workers engage with AI outputs (Leavitt et al., 2025; Lebovitz et al., 2022; Raisch & Krakowski, 2021), the cumulative-translation view reads them as carrying information about embedding and configuration alongside information about user engagement. The analytic action moves upstream to interfaces where values work is already underway before any visible outcome arrives, and toward the override patterns that offer scholars an evidentiary entry point to these interfaces.

This insight extends scholarship on conjoined agency from action to values (Humberd & Latham, 2026; Murray et al., 2021; Raisch & Krakowski, 2021). That scholarship theorizes how humans and AI jointly accomplish tasks through configurations of automation and augmentation, blurring the boundaries between human and machine contributions as their relationship shifts from coexistence toward co-creation (Waardenburg & Huysman, 2022). Our model reveals that conjoined agency also shapes what values are enacted, often without any party fully controlling, or even recognizing, the resulting prioritization. Configuration choices constrain which values AI agents can enact; enacted outcomes inform subsequent configuration; and through this recursion, responsibility for value prioritization becomes distributed across the chain as an accumulation of many localized decisions. The implication is that managing values in AI agent systems becomes a question of how organizations sustain attention across the full sequence of transformation

moments. The design choices that technical approaches emphasize, the use patterns that organizational approaches foreground, and the interfaces between them where translation work occurs each contribute to what AI agents finally enact.

Reflexivity, Improvisation, and Values Work

Our paper extends values work scholarship by theorizing how AI agents participate in values work. This line of research conceptualizes values as ongoing accomplishments achieved through practices that articulate, embody, and circulate what matters (Gehman et al., 2013). Running quietly through this work is the assumption that those who enact and circulate values also interpret them, noticing when articulated commitments no longer fit emerging circumstances and adjust enactment accordingly. This assumption holds across otherwise divergent traditions, including processual accounts that foreground practices (Gehman et al., 2013), institutional accounts that cast values as organizational commitments (Kraatz & Flores, 2015), and cognitive accounts that treat them as mental structures guiding behavior (Rokeach, 1973; Schwartz, 1992). Across these traditions, the capacity to act on values and the capacity to recognize when those values need adjustment remain coupled, even if unevenly distributed.

AI agents complicate this coupling. They can enact and circulate prioritized values at scale through patterned action while lacking the interpretive capacity to recognize when those priorities no longer hold. This incapacity is easily overlooked, because the autonomy and opacity of LLM-based agents heighten perceptions of their agency and lead people to credit them with intentionality and understanding they in fact lack (Vanneste & Puranam, 2025). Values can be performed consistently in circumstances that would otherwise invite reconsideration and adjustment, paralleling concerns that substituting machine learning for human decision-making may reduce routine diversity and exacerbate learning myopia (Balasubramanian et al., 2022).

Our model adds a values-specific account of asymmetric capacity, where articulation, enactment, and circulation become distributed across actors who differ fundamentally in their ability to recognize and respond to misfit (Murray et al., 2021; Shrestha et al., 2021). We do not presume that human reflexivity is either uniform or assured. Organizational structures, professional norms, routinization, and power all condition whose interpretations matter and when they are leveraged (Kraatz et al., 2020). The shift is that values scholarship cannot assume that the actors performing values could, in principle, also interpret when those values misfit. Values work in contexts involving AI agents, then, becomes a question of how interpretation and enactment get coupled across actors with structurally different capacities, redirecting analytical attention from how reflexivity is unevenly distributed across human actors to how it can be reconnected across actors who do not share it.

This redirection adds to ongoing conversations about human-AI complementarity, which discuss how human and AI capabilities offset one another's limitations and how alignment requires mechanisms that are dynamically scaffolded over time (Humberd & Latham, 2026; Raisch & Krakowski, 2021; Shrestha et al., 2021). Values work poses a distinctive complementarity problem. Interpretation must inform enactment, yet the two capacities locate in different actors who cannot directly communicate about misfit, and neither better specification nor tighter monitoring resolves the gap because AI agents cannot recognize when circumstances call for departure from established patterns. The conversation thus moves toward how organizations can design coupling mechanisms that reconnect interpretation to enactment across the chain of translations. In this regard, organizational forms and routines become candidate sites for theorizing distributed reflexivity. Managers, for instance, can serve as one such mechanism, reconnecting interpretation to enactment by translating an algorithmic system's logic and

correcting its outputs from within and around it, supplying value-based tradeoffs the system cannot make on its own (Leavitt et al., 2025). Among these candidate sites, we identify improvisation as an analytical priority. When workers override or work around AI agents, organizational members supply the reflexive judgment that the AI agents cannot, reasserting commitments that algorithmic enactment has missed. Such departures are familiar from accounts of how professionals weigh algorithmic recommendations against their own judgment (Lebovitz et al., 2022; Mayer et al., 2021; Raisch & Krakowski, 2021). In our account, however, improvisation also does something more fundamental. It is where interpretation re-enters the chain and where organizations can begin value inquiry as a discovery-oriented process through which they construct shared notions of the good when troublesome situations arise (Espedal & Carlsen, 2024). AI agent settings generate such situations routinely as the AI agents cannot themselves participate in the inquiry their enactments provoke. Accordingly, the conversation that follows might ask how value inquiry holds together when one of the actors who provoked it cannot participate in working out what the situation called for. This opens theoretical questions about distributed inquiry that values scholarship has not had to answer in human-focused settings.

Generative Directions for Organizational Theory

We developed this framework to explain how values transform across AI agent systems. Yet its central analytical moves grounded in assemblage theory—that is, tracing a property across stacked assemblages, locating transformation at their interfaces, and following the recursive feedback among them—are not specific to values. In this section we develop three directions in which the framework advances a conversation already under way, each reaching beyond the context in which we built it.

First, our stacked assemblage framework offers a way of tracing values across the layered infrastructures of AI agent systems, and that analytical move generalizes beyond values. The framework directs attention to how organizational properties transform across multi-layered technical infrastructures, each layer carrying its own logic and contributing differently to the property under study. This opens conversations about how legitimacy, identity, accountability, and expertise (Faraj et al., 2018; Faraj et al., 2025; Kraatz et al., 2020) move through the layered AI architectures we examine here, raising the broader question of whether organizational properties of all kinds undergo a structurally similar transformation as they move through them.

Second, our model invites scholars to reframe AI alignment from a technical specification problem to an ongoing organizational accomplishment. Technical AI literatures conceptualize alignment as ensuring AI behavior matches human values, with research focused on training-time techniques and runtime constraints (Bai et al., 2022; Christiano et al., 2017; Dong et al., 2024). Our model relocates alignment to the recursive interactions across configuration, embedding, and performance, recasting it as an ongoing accomplishment maintained through how organizations interpret, monitor, and adjust the values their AI agents enact in practice (see also Teodorescu et al., 2021, for a parallel argument in the context of fairness in human-ML augmentation). This reframing opens questions about what organizational capabilities enable continuous alignment work, how organizations recognize misalignment when the AI agent that enacts misaligned values cannot itself flag the misalignment, and under what conditions recursive cycles produce alignment that strengthens over time versus drift, pointing toward an organizational science of alignment that complements technical approaches.

Third, our analysis surfaces asymmetric capacity in values work. The asymmetry is not that AI agents cannot interpret, whereas humans can. AI agents interpret continuously, reading

context, weighing inputs, and applying configured values to situations their designers never anticipated. What they cannot do is judge those configured values themselves, recognizing when the values they enact no longer fit a situation and calling for their consideration. That reflective judgment is the capacity at issue, and the dynamic is not specific to values. Across its variants, agency scholarship has treated enactment and this reflexive judgment as capacities that travel together, whether they reside in the same purposive actor or are recombined among the human and nonhuman elements of an assemblage. Even theories of distributed and conjoined agency (Glaser et al., 2024; Humberd & Latham, 2026; Murray et al., 2021), which loosen the tie between enactment and reflexive judgment, assume that it can be supplied somewhere in the configuration. AI agents unsettle this assumption. Because they bring fluent interpretation of the cases before them but not reflexive judgment about the values they apply, the judgment that would recognize value misfit no longer co-resides with enactment in the acting agent and must be recombined deliberately, across actors with fundamentally different capacities. Theorizing asymmetric capacity opens questions about how organizations design coupling mechanisms across such actors, what new organizational forms emerge when reflexive and enactive capacities are deliberately separated, whether that separation is a durable feature of AI agents or a transitional artifact of current architectures, and whether AI agent settings might serve as a productive theoretical site for renewing organizational agency theory more broadly.

Practical Insights for Managing Values Transformation

Organizations invest substantially in AI ethics through technical solutions, governance frameworks, and compliance measures, yet continue to experience unintended values-related consequences (Lindebaum et al., 2020; Lindebaum et al., 2023). Our model suggests why these consequences persist. Values drift is an assemblage phenomenon to be navigated more than a

technical problem to be solved (Lindebaum et al., 2022). Recognizing transformation as inherent to this process reorients organizational attention from prevention to navigation, from seeking to lock values in place to developing capabilities for continuous observation and adjustment.

This reorientation carries urgency, as the time to develop these capabilities is before AI systems become more deeply embedded and autonomous and not after misalignment surfaces in consequential failures. We add that within-firm practices of this kind also lay groundwork for broader governance frameworks that demands (Humberd & Latham, 2026). Such frameworks extend beyond the firm to institutional custodians who tame disruptive algorithms by constructing envelopes or bundles of constraints that hold algorithmic behavior within boundaries a field can accept (Marti et al., 2024). Situated AI theory offers a useful starting point, identifying "recasting" (i.e., the continual adaptation of algorithms and surrounding routines) as essential to maintaining fit between AI capabilities and organizational contexts (Kemp, 2024, p. 626). Our model suggests that recasting must attend explicitly to values (Moser et al., 2024). Practically, this means configuring mechanisms to detect how espoused values are interpreted, attending to axiological encounters when AI agents meet existing routines, and systematically analyzing improvised overrides to identify where configured values fail to address operational realities. Importantly, this feedback operates bidirectionally. Human judgment provides correctives when AI enactment diverges from organizational commitments, and patterns of AI agent performance can also surface blind spots in how values were originally specified (Townsend et al., 2025). Organizations that recognize these recursive dynamics, and that integrate technical and organizational perspectives in doing so, are better positioned to manage values transformation than those treating configuration as a one-time design problem.

Boundary Conditions and an Empirical Agenda

Our claims carry boundary conditions that also mark where the framework can be taken next. First, we focused on LLM-based AI agents, yet transformation dynamics may differ for other AI architectures, such as computer vision systems, reinforcement learning agents, or hybrid systems (Jarrahi & Glaser, 2025), and emerging calls to distinguish automation from iteration in AI task design (Townsend et al., 2025) may map onto different dynamics across our three moments. Second, we bounded our analysis within organizations, yet values transformation likely propagates across supply chains, platform ecosystems, and industries, and extending the analysis across these levels could enrich the framework. Third, we developed the model through theoretical integration and illustrative cases rather than systematic empirical investigation.

Systematic empirical research is the most direct next step. Tracing values transformation through complete cycles would require longitudinal research capable of documenting how specific values evolve across assemblage boundaries (Glaser et al., 2021; Hyysalo et al., 2025). Ethnographic observation paired with computational analysis of AI agent operations could illuminate the micro-processes through which values transform at the interfaces between technical configuration and organizational practice. Specific questions worth empirical attention include how technical teams recognize when inherited axiological substrates conflict with organizational commitments; how organizations engage stakeholders in negotiating commensuration trade-offs when values are culturally situated; which organizational capabilities enable managers to interpret enacted values from performance patterns; and under what conditions human improvisation leads to reconfiguration rather than dismissal as noncompliance. Researchers might also investigate how AI agents themselves can be used reflexively to question

configured parameters when enacted outcomes diverge from organizational commitments, and what conditions enable such use.

CONCLUSION

As AI agents increasingly mediate consequential organizational functions, the familiar gap between espoused and enacted values (Argyris & Schön, 1974; Bourne & Jenkins, 2013; Schein, 2010) takes on renewed urgency. Our process model specifies how AI agents change the nature of this divergence by technologically mediating values transformation and distributing the process across assemblage interfaces: inherited orientations shape what can be configured; configured parameters meet routines with their own enacted histories; and embedded patterns encounter situated practice in ways that generate gaps that specification alone does not eliminate. By tracing how espoused values move across the LLM, AI agent, and algorithmic routine assemblages, we have shown how, where, and through what mechanisms transformation occurs.

REFERENCES

Abbott, A. (1988). *The System of Professions: An Essay on the Division of Expert Labor* (1st ed.). University of Chicago Press.

Agrawal, A., Gans, J., & Goldfarb, A. (2022). *Prediction Machines, Updated and Expanded: The Simple Economics of Artificial Intelligence* (Revised edition). Harvard Business Review Press.

Akrich, M. (1992). The De-Scription of Technical Objects. In W. E. Bijker & J. Law (Eds.), *Shaping Technology/ Building Society* (pp. 205–224). MIT Press.

Alaimo, C., & Kallinikos, J. (2021). Managing by Data: Algorithmic Categories and Organizing. *Organization Studies*, *42*(9), 1385–1407. https://doi.org/10.1177/0170840620934062

Alaimo, C., & Kallinikos, J. (2022). Organizations Decentered: Data Objects, Technology and Knowledge. *Organization Science*, *33*(1), 19–37. https://doi.org/10.1287/orsc.2021.1552

Alaimo, C., & Kallinikos, J. (2024). *Data Rules: Reinventing the Market Economy*. The MIT Press.

Anthony, C., Bechky, B. A., & Fayard, A.-L. (2023). "Collaborating" with AI: Taking a System View to Explore the Future of Work. *Organization Science*, *34*(5), 1672–1694. https://doi.org/10.1287/orsc.2022.1651

Argyris, C., & Schön, D. A. (1974). *Theory in Practice: Increasing Professional Effectiveness*. Jossey-Bass.

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T., Clark, J., … Kaplan, J. (2021). *A General Language Assistant as a Laboratory for Alignment* (No. arXiv:2112.00861). arXiv. https://doi.org/10.48550/arXiv.2112.00861

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., … Kaplan, J. (2022). *Constitutional AI: Harmlessness from AI Feedback* (No. arXiv:2212.08073). arXiv. https://doi.org/10.48550/arXiv.2212.08073

Bailey, D. E., & Barley, S. R. (2020). Beyond design and use: How scholars should study intelligent technologies. *Information and Organization*, *30*, 100286. https://doi.org/10.1016/j.infoandorg.2019.100286

Balasubramanian, N., Ye, Y., & Xu, M. (2022). Substituting Human Decision-Making with Machine Learning: Implications for Organizational Learning. *Academy of Management Review, 47*(3), 448–65. https://doi.org/10.5465/amr.2019.0470

Bankins, S., & Formosa, P. (2023). The ethical implications of artificial intelligence (AI) for meaningful work. *Journal of Business Ethics*, *185*(4), 725-740.

Beane, M. (2024). *The Skill Code: How to Save Human Ability in an Age of Intelligent Machines*. HarperCollins.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Berti, L., Giorgi, F., & Kasneci, G. (2025). *Emergent Abilities in Large Language Models: A Survey* (No. arXiv:2503.05788; Ver.2). arXiv. https://doi.org/10.48550/arXiv.2503.05788

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S. von, Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). *On the Opportunities and Risks of Foundation Models* (No. arXiv:2108.07258). arXiv. https://doi.org/10.48550/arXiv.2108.07258

Bowker, G. C., & Star, S. L. (2000). *Sorting Things Out: Classification and Its Consequences*. The MIT Press.

Bourne, H., & Jenkins, M. (2013). Organizational Values: A Dynamic Perspective. *Organization Studies*. 34(4), 495-514.

Brayne, S. (2017). Big Data Surveillance: The Case of Policing. *American Sociological Review*, *82*(5), 977–1008. https://doi.org/10.1177/0003122417725865

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512. https://doi.org/10.1177/2053951715622512

Cabantous, L., Gond, J.-P., & Johnson-Cramer, M. (2010). Decision Theory as Practice: Crafting Rationality in Organizations. *Organization Studies*, *31*(11), 1531–1566. https://doi.org/10.1177/0170840610380804

Cameron, L. (2024). The Making of the "Good Bad" Job: How Algorithmic Management Manufactures Consent through Constant and Confined Choices. *Administrative Science Quarterly, 69*(2), 458-514. https://doi.org/10.1177/00018392241236163

Christian, B. (2020). *The Alignment Problem: Machine Learning and Human Values*. W. W. Norton & Company.

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). *Deep reinforcement learning from human preferences* (No. arXiv:1706.03741). arXiv. https://doi.org/10.48550/arXiv.1706.03741

Christin, A. (2020). *Metrics at Work: Journalism and the Contested Meaning of Algorithms*. Princeton University Press.

Ciriello, R., Hannon, O., Chen, A. Y., & Vaast, E. (2024). *Ethical Tensions in Human-AI Companionship: A Dialectical Inquiry into Replika*. https://hdl.handle.net/10125/106433

Currie, R. (2024, January 23). *DPD chatbot goes off the rails at suggestion of customer*. The Register. https://www.theregister.com/2024/01/23/dpd_chatbot_goes_rogue/

D'Adderio, L. (2008). The Performativity of Routines: Theorising the Influence of Artefacts and Distributed Agencies on Routines Dynamics. *Research Policy*, *37*(5), 769–789. https://doi.org/10.1016/j.respol.2007.12.012

DeLanda, M. (2006). *A New Philosophy of Society: Assemblage Theory and Social Complexity* (Annotated edition). Continuum.

DeLanda, M. (2016). *Assemblage Theory* (1 edition). Edinburgh University Press.

Deleuze, G., & Guattari, F. (1987). *A Thousand Plateaus: Capitalism and Schizophrenia* (B. Massumi, Trans.; 2 edition). University of Minnesota Press.

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., & Gardner, M. (2021). *Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus* (No. arXiv:2104.08758). arXiv. https://doi.org/10.48550/arXiv.2104.08758

Dong, Y., Mu, R., Jin, G., Qi, Y., Hu, J., Zhao, X., Meng, J., Ruan, W., & Huang, X. (2024). *Building Guardrails for Large Language Models* (No. arXiv:2402.01822; Version 1). arXiv. https://doi.org/10.48550/arXiv.2402.01822

Elish, M. C. (2019). Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science, Technology, and Society, 5*, 40–60. https://doi.org/10.17351/ests2019.260

Espedal, G., & Carlsen, A. (2024). Value inquiry and constructing the good in organizations. *Organization Studies*, *45*(8), 1075-1098. https://doi.org/10.1177/01708406241253161

Espeland, W. N. (2001). Value-Matters. *Economic and Political Weekly*, *36*(21), 1839–1845.

Espeland, W. N., & Stevens, M. L. (1998). Commensuration as a Social Process. *Annual Review of Sociology*, *24*, 313–343.

Faraj, S., Pachidi, S., & Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. *Information and Organization*, *28*(1), 62–70. https://doi.org/10.1016/j.infoandorg.2018.02.005

Faraj, S., Perez-Torrents, J., & Bhardwaj, A. (2025). A Time for Monsters: Organizational Knowing After Large Language Models. *Strategic Organization, 24*(2), 343-356. https://doi.org/10.1177/14761270251410675

Feldman, M. S., & Pentland, B. T. (2003). Reconceptualizing organizational routines as a source of flexibility and change. *Administrative Science Quarterly*, *48*(1), 94–118.

Følstad, A., Nordheim, C., & Bjørkli, C. (2018). *What Makes Users Trust a Chatbot for Customer Service? An Exploratory Interview Study* In: Bodrunova, S. (eds) *Internet Science.* (pp. 194–208). Springer. https://doi.org/10.1007/978-3-030-01437-7_16

Friedman, B., & Hendry, D. G. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination* (Illustrated edition). The MIT Press.

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

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). *Retrieval-Augmented Generation for Large Language Models: A Survey* (No. arXiv:2312.10997). arXiv. https://doi.org/10.48550/arXiv.2312.10997

Gehman, J., Sharma, G., & Beveridge, 'Alim. (2022). Theorizing Institutional Entrepreneuring: Arborescent and rhizomatic assembling. *Organization Studies*, *43*(2), 289–310. https://doi.org/10.1177/01708406211044893

Gehman, J., Treviño, L. K., & Garud, R. (2013). Values Work: A Process Study of the Emergence and Performance of Organizational Values Practices. *Academy of Management Journal*, *56*(1), 84–112.

Glaser, V. L. (2014). Enchanted Algorithms: The Quantification of Organizational Decision-Making. *Doctoral dissertation, University of Southern California, Marshall School of Business*.

Glaser, V. L. (2017). Design Performances: How Organizations Inscribe Artifacts to Change Routines. *Academy of Management Journal*, *60*(6), 2126–2154. https://doi.org/10.5465/amj.2014.0842

Glaser, V. L., Pollock, N., & D'Adderio, L. (2021). The Biography of an Algorithm: Performing Algorithmic Technologies in Organizations. *Organization Theory*, *2*, 1–27. https://doi.org/10.1177/26317877211004609

Glaser, V. L., Sloan, J., & Gehman, J. (2024). Organizations as Algorithms: A New Metaphor for Advancing Management Theory. *Journal of Management Studies*, *61*(6), 2748–2769. https://doi.org/10.1111/joms.13033

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 2053951719897945. https://doi.org/10.1177/2053951719897945

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). *Measuring Massive Multitask Language Understanding* (No. arXiv:2009.03300). arXiv. https://doi.org/10.48550/arXiv.2009.03300

Henfridsson, O. & Bygstad, B. (2013). The Generative Mechanisms of Digital Infrastructure Evolution. *MIS Quarterly, 37*(3), 907-931. https://doi.org/10.25300/MISQ/2013/37.3.11

Hillebrand, L., Raisch, S., & Schad, J. (2025). Managing with Artificial Intelligence: An Integrative Framework. *Academy of Management Annals*, *19*(1), 343–375. https://doi.org/10.5465/annals.2022.0072

Hinings, B., Gegenhuber, T., & Greenwood, R. (2018). Digital innovation and transformation: An institutional perspective. *Information and Organization*, *28*(1), 52–61. https://doi.org/10.1016/j.infoandorg.2018.02.004

Hosanagar, K., & Krishnan, R. (2024). Who Profits the Most from Generative AI? *MIT Sloan Management Review*, *65*(3), 24–29.

Howard-Grenville, J. (2005). The Persistence of Flexible Organizational Routines: The Role of Agency and Organizational Context. *Organization Science*, *16*(6), 618–636. https://doi.org/10.2307/25146000

Howell, A., Kirk-Brown, A., & Cooper, B. K. (2012). Does congruence between espoused and enacted organizational values predict affective commitment in Australian organizations? *The International Journal of Human Resource Management*, *23*(4), 731–747. https://doi.org/10.1080/09585192.2011.561251

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

Hyysalo, S., Pollock, N., Seibt, D., & Williams, R. (2025). Biographies of artifacts and practices. In I. Schulz-Schaeffer, A. Windeler, & B. Blättel-Mink (Eds.), *Handbook of Innovation*. Springer, Cham. https://doi.org/10.1007/978-3-031-25143-6_48-1

Jarrahi, M. H., & Glaser, V. L. (2025). Not all AI systems are created equal: A typology of AI systems' performance and affordance. In *Proceedings of the 58th Hawaii International Conference on System Sciences (HICSS)*. https://doi.org/10.24251/HICSS.2025.655

Kallinikos, J., Aaltonen, A., & Marton, A. (2013). The Ambivalent Ontology of Digital Artifacts. *MIS Quarterly*, *37*(2), 357–370.

Kellogg, K., Valentine, M., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. *Academy of Management Annals*, *14*(1), 366–410. https://doi.org/10.5465/annals.2018.0174

Kemp, A. (2024). Competitive Advantage Through Artificial Intelligence: Toward a Theory of Situated AI. *Academy of Management Review, 49*(3), 618–35. https://doi.org/10.5465/amr.2020.0205

Kim, B., Rezazade Mehrizi, M., & Bailey, D. (2025). Interlacing Situated and Algorithmic Modes of Knowledge Work: A Workplace Jurisdiction Perspective. *Academy of Management Journal, 68*(5), 907–38. https://doi.org/10.5465/amj.2023.1286.

Kraatz, M. S., & Block, E. S. (2008). Organizational Implications of Institutional Pluralism. In R. Greenwood, C. Oliver, R. Suddaby, & K. Sahlin-Andersson (Eds.), *The SAGE*

Handbook of Organizational Institutionalism (pp. 243‑275). SAGE Publications Ltd. https://doi.org/10.4135/9781849200387.n10

Kraatz, M. S., & Flores, R. (2015). Introduction. In *Institutions and Ideals: Philip Selznick's Legacy for Organizational Studies* (Vol. 44, pp. 1–7). Emerald Group Publishing Limited. https://doi.org/10.1108/S0733-558X20150000044001

Kraatz, M. S., Flores, R., & Chandler, D. (2020). The Value of Values for Institutional Analysis. *Academy of Management Annals*, *14*(2), 474–512. https://doi.org/10.5465/annals.2018.0074

La Cava, W. G. (2023). Optimizing fairness tradeoffs in machine learning with multiobjective meta-models. *Proceedings of the Genetic and Evolutionary Computation Conference*, 511–519. https://doi.org/10.1145/3583131.3590487

Leavitt, K., Barnes, C. M., & Shapiro, D. L. (2025). The role of human managers within algorithmic performance management systems: A process model of employee trust in managers through reflexivity. *Academy of Management Review, 50*(4), 745-767. https://doi.org/10.5465/amr.2022.0058

Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. (2022). To Engage or Not to Engage with AI for Critical Judgments: How Professionals Deal with Opacity When Using AI for Medical Diagnosis. *Organization Science*, *33*(1), 126–148. https://doi.org/10.1287/orsc.2021.1549

Leonardi, P. M. (2011). When Flexible Routines Meet Flexible Technologies: Affordance, Constraint, and the Imbrication of Human and Material Agencies. *MIS Quarterly*, *35*(1), 147–168.

Lewis, K., Lange, D., & Gillis, L. (2005). Transactive Memory Systems, Learning, and Learning Transfer. *Organization Science*, *16*(6), 581–598.

Lindebaum, D., Vesa, M., & den Hond, F. (2020). Insights from "The Machine Stops" to Better Understand Rational Assumptions in Algorithmic Decision Making and its Implications for Organizations. *Academy of Management Review, 45*(1), 247–263. https://doi.org/10.5465/amr.2018.0181

Lindebaum, D., Glaser, V. L., Moser, C., & Ashraf, M. (2022). When Algorithms Rule, Values Can Wither. *MIT Sloan Management Review*, *64*(2), 66–69.

Lindebaum, D., Moser, C., Ashraf, M., & Glaser, V. L. (2023). Reading the technological society to understand the mechanization of values and its ontological consequences. *Academy of Management Review*, *48*(3), 575–592. https://doi.org/10.5465/amr.2021.0159

Lindebaum, D., Moser, C., & Islam, G. (2024). Big Data, Proxies, Algorithmic Decision-Making and the Future of Management Theory. *Journal of Management Studies*, *61*(6), 2724–2747. https://doi.org/10.1111/joms.13032

Luo, J., Zhang, W., Yuan, Y., Zhao, Y., Yang, J., Gu, Y., Wu, B., Chen, B., Qiao, Z., Long, Q., Tu, R., Luo, X., Ju, W., Xiao, Z., Wang, Y., Xiao, M., Liu, C., Yuan, J., Zhang, S., … Zhang, M. (2025). *Large Language Model Agent: A Survey on Methodology, Applications and Challenges* (No. arXiv:2503.21460). https://doi.org/10.48550/arXiv.2503.21460

Marti, E., Lawrence, T., & Steele, C. (2024). Constructing envelopes: How institutional custodians can tame disruptive algorithms. *Academy of Management Journal*, *67*(5), 1273–1301.

Mayer, A. S., van den Broek, E., & Karačić, T. (2026). 'Let Me Explain': A Comparative Field Study on How Experts Enact Authority Over Clients When Facing AI Decisions. *Journal of Management Studies*, *63*(2), 366-398.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). *A Survey on Bias and Fairness in Machine Learning* (No. arXiv:1908.09635). arXiv. https://doi.org/10.48550/arXiv.1908.09635

Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., Grave, E., LeCun, Y., & Scialom, T. (2023). *Augmented Language Models: A Survey* (No. arXiv:2302.07842). arXiv. https://doi.org/10.48550/arXiv.2302.07842

Moffatt v. Air Canada, No. 149 (British Columbia Civil Resolution Tribunal 2024). https://www.canlii.org/en/bc/bccrt/doc/2024/2024bccrt149/2024bccrt149.html

Möhlmann, M., Zalmanson, L., Henfridsson, O., & Gregory, R.W. (2021). Algorithmic Management of Work on Online Labor Platforms: When Matching Meets Control. *MIS Quarterly, 45*(4), 1999-2022. https://doi.org/10.25300/MISQ/2021/15333

Monteiro, E. (2022). *Digital Oil: Machineries of Knowing*. The MIT Press.

Moser, C., Glaser, V., & Lindebaum, D. (2024). Taking Situatedness Seriously in Theorizing About Competitive Advantage Through Artificial Intelligence: A Response to Kemp's 'Competitive Advantages Through Artificial Intelligence.' *Academy of Management Review, 49*(3), 683–85. https://doi.org/10.5465/amr.2023.0265

Murray, A., Rhymer, J., & Sirmon, D. (2021). Humans and Technology: Forms of Conjoined Agency in Organizations. *Academy of Management Review*, *46*(3), 552–571.

Omidvar, O., Safavi, M., & Glaser, V. L. (2023). Algorithmic Routines and Dynamic Inertia: How Organizations Avoid Adapting to Changes in the Environment. *Journal of Management Studies*, *60*(2), 313–345. https://doi.org/10.1111/joms.12819

Orlikowski, W. J., & Scott, S. V. (2008). Sociomateriality: Challenging the Separation of Technology, Work and Organization. *The Academy of Management Annals*, *2*(1), 433–474. https://doi.org/10.1080/19416520802211644

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). *Training language models to follow instructions with human feedback* (No. arXiv:2203.02155). arXiv. https://doi.org/10.48550/arXiv.2203.02155

Pachidi, S., Berends, H., Faraj, S., & Huysman, M. (2021). Make Way for the Algorithms: Symbolic Actions and Change in a Regime of Knowing. *Organization Science*, *32*(1), 18–41. https://doi.org/10.1287/orsc.2020.1377

Pentland, B. T., & Feldman, M. S. (2008). Designing Routines: On the Folly of Designing Artifacts, While Hoping for Patterns of Action. *Information and Organization*, *18*(4), 235–250. https://doi.org/10.1016/j.infoandorg.2008.08.001

Pentland, B. T., & Rueter, H. H. (1994). Organizational Routines as Grammars of Action. *Administrative Science Quarterly*, *39*(3), 484–510. https://doi.org/10.2307/2393300

Pollock, N., & Williams, R. (2009). *Software and organisations: The biography of the enterprise-wide system or how SAP conquered the world*. Routledge, 2009.

Porter, T. M. (1996). *Trust in Numbers*. Princeton University Press.

Power, M. (1997). *The Audit Society: Rituals of Verification.* Oxford University Press.

Rahwan, I. (2018). Society-in-the-Loop: Programming the Algorithmic Social Contract. *Ethics and Information Technology*, *20*(1), 5–14. https://doi.org/10.1007/s10676-017-9430-8

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

Raisch, S., & Fomina, K. (2025). Combining Human and Artificial Intelligence: Hybrid Problem-Solving In Organizations. *Academy of Management Review*, *50*(2), 441-464.

Rokeach, M. (1973). *The Nature of Human Values*. Free Press.

Saifer, A., & Dacin, M. T. (2022). Data and Organization Studies: Aesthetics, emotions, discourse and our everyday encounters with data. *Organization Studies*, *43*(4), 623–636. https://doi.org/10.1177/01708406211006250

Salvato, C., & Rerup, C. (2018). Routine Regulation: Balancing Conflicting Goals in Organizational Routines. *Administrative Science Quarterly*, *63*(1), 170–209. https://doi.org/10.1177/0001839217707738

Sapkota, R., Roumeliotis, K. I., & Karkee, M. (2025). *AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges* (No. arXiv:2505.10468). https://doi.org/10.48550/arXiv.2505.10468

Schein, E. H. (2010). *Organizational Culture and Leadership* (4 edition). Jossey-Bass.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). *Toolformer: Language Models Can Teach Themselves to Use Tools* (No. arXiv:2302.04761). arXiv. https://doi.org/10.48550/arXiv.2302.04761

Schwartz, S. H. (1992). Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 25, pp. 1–65). Academic Press. https://doi.org/10.1016/S0065-2601(08)60281-6

Selznick, P. (1957). *Leadership in Administration: A Sociological Interpretation*. Harper and Row.

Shneiderman, B. (2022). *Human-centered AI*. Oxford University Press. https://doi.org/10.1093/oso/9780192845290.001.0001

Shrestha, Y. R., Ben-Menahem, S. M., & von Krogh, G. (2019). Organizational Decision-Making Structures in the Age of Artificial Intelligence. *California Management Review*, *61*(4), 66–83. https://doi.org/10.1177/0008125619862257

Shrestha, Y. R., Krishna, V., & von Krogh, G. (2021). Augmenting organizational decision-making with deep learning algorithms: Principles, promises, and challenges. *Journal of Business Research*, *123*, 588-603.

Stark, D. (2009). *The Sense of Dissonance: Accounts of Worth in Economic Life*. Princeton University Press.

Stark, D., & van den Broeck, P. E. (2024). Principles of Algorithmic Management. *Organization Theory*, *5*(2), 26317877241257213. https://doi.org/10.1177/26317877241257213

Stelmaszak, M., Joshi, M., & Constantiou, I. (2026). Artificial Intelligence as an Organizing Capability Arising from Human-Algorithm Relations. *Journal of Management Studies*, *63*(2). https://doi.org/10.1111/joms.70003

Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy*, *42*(9), 1568–1580. https://doi.org/10.1016/j.respol.2013.05.008

Suchman, L. (2007). *Human-Machine Reconfigurations: Plans and Situated Actions* (2 edition). Cambridge University Press.

Teodorescu, M. H., Morse, L., Awwad, Y., & Kane, G. C. (2021). Failures of fairness in automation require a deeper understanding of human–ml augmentation. *MIS Quarterly*, *45*(3), 1483-1499. https://doi.org/10.25300/MISQ/2021/16535

Tilson, D., Lyytinen, K., Sørensen, K. (2010). Digital Infrastructures: The Missing IS Research Agenda. *Information Systems Research, 21*(4), 748-759. https://doi.org/10.1287/isre.1100.0318

Townsend, D. M., Hunt, R. A., Rady, J., Manocha, P., & Jin, J. H. (2025). Are the futures computable? Knightian uncertainty and artificial intelligence. *Academy of Management Review*, *50*(2): 415‑440. https://doi.org/10.5465/amr.2022.0237

Tsoukas, H., & Chia, R. (2002). On Organizational Becoming: Rethinking Organizational Change. *Organization Science*, *13*(5), 567–582.

Turner, S. F., & Rindova, V. (2012). A Balancing Act: How Organizations Pursue Consistency in Routine Functioning in the Face of Ongoing Change. *Organization Science*, *23*(1), 24–46. https://doi.org/10.1287/orsc.1110.0653

van Bommel, K., Rasche, A., & Spicer, A. (2023). From Values to Value: The Commensuration of Sustainability Reporting and the Crowding Out of Morality. *Organization & Environment*, *36*(1), 179–206. https://doi.org/10.1177/10860266221086617

Vanneste, B., & Puranam, P. (2025). Artificial intelligence, trust, and perceptions of agency. *Academy of Management Review, 50*(4), 726‑744. https://doi.org/10.5465/amr.2022.0041

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (No. arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

von Krogh, G. (2018). Artificial Intelligence in Organizations: New Opportunities for Phenomenon-Based Theorizing. *Academy of Management Discoveries*, *4*(4), 404–409. https://doi.org/10.5465/amd.2018.0084

Waardenburg, L., & Huysman, M. (2022). From Coexistence to Co-Creation: Blurring Boundaries in the Age of AI. *Information and Organization, 32*(4), 1–11. https://doi.org/10.1016/j.infoandorg.2022.100432

Waardenburg, L., Huysman, M., & Sergeeva, A. (2022). In the Land of the Blind, the One-Eyed Man Is King: Knowledge Brokerage in the Age of Learning Algorithms. *Organization Science, 33*(1), 59–82. https://doi.org/10.1287/orsc.2021.1544

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong* (pp. xi, 275). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195374049.001.0001

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

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2023). *A Survey of Large Language Models* (No. arXiv:2303.18223). arXiv. https://doi.org/10.48550/arXiv.2303.18223

FIGURE 1 — A PROCESS MODEL OF VALUES TRANSFORMATION IN AI AGENT ASSEMBLAGES

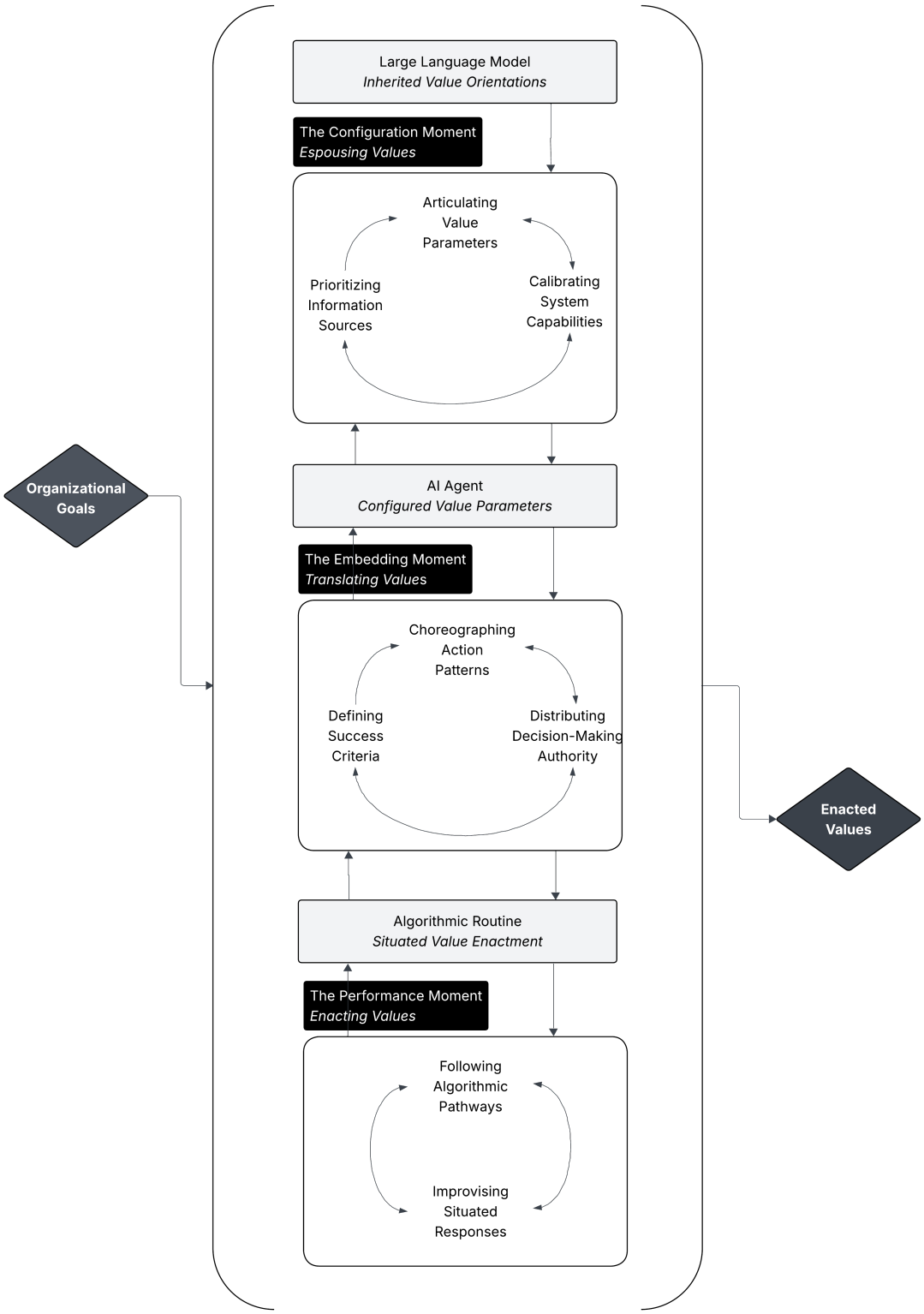

TABLE 1 — COMPARISON OF THE THREE ASSEMBLAGES

Dimension	LLM Assemblage	AI Agent Assemblage	Algorithmic Routine Assemblage
Primary Focus	Foundational capabilities	Configured behavior	Situated practice
Axiological Character	Inherited value orientations (axiological substrate)	Configured value parameters	Situated value enactment
Representative Elements	Training data, neural architecture, optimization procedures	System prompts, guardrails, memory systems, tool integrations	Workflows, metrics, authority structures, human-AI interactions
Espoused Values	Pre-espoused values inherited through design and training	Explicit configuration choices by technical teams and organizational leaders	Carried forward from configuration; translated into routines
Enacted Values	Inherited biases, reasoning patterns, implicit hierarchies	AI interpretation of configured values	Visible behavior in practice; gaps become observable
Human Agency	Indirect: shaped at deployment by prior data curators, trainers, and architecture designers	Technical teams and organizational leaders configure	Managers observe, intervene, reconfigure; workers improvise
AI Agency	Training process determines orientations	Interprets configured values	Performs, follows algorithmic pathways
Key Transformation	Pre-existing orientations become available for prioritization	Inherited → Explicit (espousing values)	Designed → Practiced (enacting values)

AUTHOR BIOGRAPHIES

Vern L. Glaser is a professor of entrepreneurship and family enterprise in the Department of Strategy, Entrepreneurship and Management at the Alberta School of Business, University of Alberta, Canada, and academic director of the Alberta Business Family Institute. He received his PhD from the University of Southern California. His research examines algorithmic organizing, the role of AI and data analytics in decision-making and governance, and family enterprise.

Jennifer Sloan is a research fellow at the UCL School of Management. Her research examines algorithmic organizing and focuses on how AI and analytics impact what organizations perceive, decide, and do. Jennifer holds a BEd and an MBA from the University of Alberta and a PhD in Strategic Management and Organization from the University of Alberta. Prior to her academic career, she worked in management consulting and education.

Rodrigo Valadão is a global executive serving as Head of Business Development at Porto do Açu, Latin America's largest private port-industrial complex and a leading hub for the energy transition in Brazil. He holds a PhD in Strategy, Entrepreneurship, and Management from the University of Alberta and is currently a postdoctoral researcher at FGV EBAPE. His research examines the intersection of AI and analytics, strategy, innovation, and organizing. Rodrigo brings nearly two decades of executive experience across the Americas and Europe and serves as a council member and advisor in the oil, gas, and energy sectors.

Evelyn R. Micelotta is an Associate Professor and holder of the Steven Grossman Endowed Chair in Family Business at the University of Vermont’s Grossman School of Business. She received a PhD in Strategic Management and Organization from the Alberta School of Business (University of Alberta, Canada) and a PhD in Management Engineering from Milan Polytechnic. Her research explores institutional processes and cultural dynamics in organizations.