

DOI: https://doi.org/10.48009/3_iis_2026_126

From AI ethics to ethical misalignment: understanding emerging risks in agentic AI systems

Dr. Deb Tech, *Dakota State University*, deb.tech@dsu.edu

Dr. Aravindh Sekar, *Dakota State University*, aravindh.sekar@trojans.dsu.edu

Abstract

The emergence of Agentic AI systems is reshaping organizational processes through continuous adaptation, autonomous coordination, and real-time operational execution. Unlike traditional artificial intelligence applications focused on discrete decision-making tasks, Agentic AI systems operate continuously and interact dynamically with organizational environments. Existing AI ethics literature has primarily emphasized issues such as fairness, bias, transparency, and accountability within static or decision-centric AI contexts. Similarly, traditional AI alignment research has largely focused on ensuring that AI systems pursue intended goals and remain controllable under specified operational conditions. However, these approaches may be insufficient for explaining how ethical risks evolve with continuously operating autonomous systems. This paper develops a conceptual framework for understanding ethical misalignment in Agentic AI systems. Ethical misalignment is defined as the gradual deviation between intended ethical objectives and emergent system outcomes resulting from ongoing adaptation and autonomous interaction. Rather than focusing primarily on goal alignment or system control, the study conceptualizes ethical misalignment as a longitudinal organizational phenomenon emerging through continuous adaptation and interaction across adaptive enterprise ecosystems. This study examines limitations of traditional AI ethics frameworks and proposes a framework linking ethical design, adaptive behavior, emergent interactions, and ethical misalignment outcomes, contributing to an emergent perspective on ethics within autonomous organizational systems.

Keywords: Agentic AI; AI Ethics; Ethical Misalignment; Autonomous Systems; Responsible AI; Information Systems

Introduction

Artificial intelligence (AI) has become increasingly embedded within organizational processes, enterprise systems, and digital infrastructures, transforming how organizations analyze information, automate activities, and coordinate operational workflows. Recent advances in AI have accelerated the emergence of Agentic AI systems capable of autonomous interaction, adaptive learning, and real-time execution across complex organizational environments. Unlike traditional AI applications that primarily support human decision making, Agentic AI systems increasingly operate as autonomous organizational participants capable of initiating actions, coordinating activities, and continuously adapting their behavior in response to changing operational conditions (Davenport & Ronanki, 2018; Brynjolfsson & McAfee, 2017).

The rapid expansion of AI adoption has also intensified concerns surrounding ethical governance, fairness, accountability, transparency, and responsible AI implementation. Existing AI ethics research has established important principles for addressing issues such as algorithmic bias, explainability, discrimination, privacy, and governance (Floridi et al., 2018; Floridi & Cows, 2019; Jobin et al., 2019). While these frameworks provide an important foundation for responsible AI, they have largely been developed around AI systems that generate identifiable outputs or support discrete decision-making processes. Emerging discussions suggest that Agentic AI systems introduce additional ethical challenges associated with continuous operation, autonomous behavior, and reduced human oversight (e.g., Hahn et al., 2026; Lynch et al., 2025).

This evolution creates an important gap in existing literature. Because Agentic AI systems continuously adapt, interact with their environments, and modify their operational behavior over time, ethical risks may emerge gradually rather than at isolated decision points. Consequently, understanding ethics in Agentic AI environments requires moving beyond static, decision-centric perspectives toward explanations that account for continuous adaptation, autonomous coordination, and evolving organizational interactions.

To address this gap, this paper develops a process-oriented framework for understanding ethical misalignment in Agentic AI systems. Building upon established AI ethics research and emerging work on autonomous AI systems, the paper introduces the concept of ethical misalignment to explain how ethical risks may emerge and evolve within continuously adaptive Agentic AI environments. Rather than viewing ethical failures as isolated events, the proposed framework positions ethical misalignment as an emergent organizational phenomenon produced through ongoing optimization and interaction

This study contributes to the Information Systems literature in three ways. First, it extends AI ethics research by proposing ethical misalignment as a process-oriented construct for understanding ethical risk in continuously adaptive AI systems. Second, it integrates complex adaptive systems and sociotechnical perspectives to explain how ethical outcomes emerge through ongoing interactions among AI systems, organizational processes, and their environments. Third, it reframes ethical governance as a dynamic organizational capability requiring continuous monitoring and adaptation rather than a static compliance mechanism established at system deployment.

Research Approach

This study adopts a conceptual research approach to examine emerging ethical risks associated with Agentic AI systems. Rather than relying on empirical data collection, the paper develops its arguments through the synthesis and integration of existing literature across AI ethics, autonomous systems, Agentic AI, and information systems research. Conceptual research is particularly appropriate for emerging technological phenomena where theoretical understanding remains underdeveloped and rapidly evolving (Jaakkola, 2020). By synthesizing literature on AI ethics, AI governance, autonomous agents, AI misalignment, multi-agent systems, and adaptive AI behavior, this study identifies limitations in existing ethical frameworks when applied to Agentic AI environments. The analysis also incorporates process-oriented and complex adaptive systems perspectives, which conceptualize organizational outcomes as emerging through continuous interaction, adaptation, and feedback among interconnected systems and actors. This perspective aligns with process-oriented traditions within Information Systems research that conceptualize organizational technologies not as static artifacts, but as evolving sociotechnical systems shaped through continuous interaction, adaptation, and enactment (Orlikowski, 2000; Van de Ven, 2007). Similarly, sociomaterial perspectives in IS emphasize that technological behavior and organizational outcomes emerge relationally through ongoing interaction between digital systems, organizational routines, and human actors rather than through isolated technological decisions alone. From this perspective, ethical

outcomes within Agentic AI environments are understood not as isolated algorithmic decisions, but as dynamic and emergent conditions shaped through ongoing autonomous coordination and evolving system interaction.

Through this conceptual synthesis, the paper identifies a critical limitation in existing AI ethics research. Existing AI ethics literature provides important foundations for responsible AI governance but offers comparatively limited explanation for how ethical risks evolve within continuously adaptive autonomous environments. However, Agentic AI systems increasingly operate as autonomous organizational actors within evolving enterprise environments. These characteristics increasingly position Agentic AI environments as complex adaptive organizational systems in which ethical outcomes emerge through distributed interaction and continuous feedback. Consequently, existing ethical frameworks may be insufficient for explaining how ethical risks emerge longitudinally through ongoing system behavior.

To address this gap, the study develops the concept of ethical misalignment as a process-oriented framework for understanding evolving ethical risks within Agentic AI systems. Rather than conceptualizing ethics as a fixed governance condition validated during system deployment, the paper positions ethical misalignment as an evolving organizational process shaped by continuous adaptation, autonomous interaction, and emergent interaction patterns across interconnected autonomous systems.

Theoretical Foundations and Ethical Challenges in Agentic AI

The emergence of Agentic AI systems challenges many assumptions underlying traditional AI ethics frameworks. Unlike conventional AI applications that generate discrete recommendations or predictions, Agentic AI systems operate within dynamic organizational environments. These characteristics fundamentally alter how ethical risks emerge and therefore require theoretical perspectives capable of explaining ethics as an evolving organizational phenomenon rather than as isolated decision events.

This shift also requires clearer conceptual differentiation between ethical misalignment and related constructs frequently discussed within AI ethics and governance literature. While ethical misalignment may intersect with concerns surrounding algorithmic bias, AI alignment, AI safety, governance failure, and operational drift, the construct developed in this study focuses specifically on the gradual organizational divergence between intended ethical objectives and emergent operational behaviors within continuously adaptive autonomous systems. Unlike algorithmic bias, which typically focuses on discriminatory outcomes within model outputs, and AI alignment research, which primarily examines whether systems pursue intended goals under specified conditions, ethical misalignment represents a broader and more dynamic construct. It refers to the longitudinal ethical divergence resulting from continuous adaptation, autonomous coordination, and interactions among distributed actors and systems within organizational environments.

To clarify the conceptual contribution of this study, Table 1 compares ethical misalignment with several closely related concepts discussed in the AI ethics and Information Systems literature. Although these concepts address important ethical concerns associated with AI systems, they differ in their primary focus, temporal orientation, and analytical perspective. This comparison demonstrates that ethical misalignment represents a distinct process-oriented construct for understanding how ethical risks emerge through continuous adaptation, autonomous interaction, and evolving organizational behavior within Agentic AI systems

Table 1. Differentiating Ethical Misalignment from Related AI Ethics Constructs

Construct	Primary Focus	Nature of Risk	Temporal Orientation	Primary Unit of Analysis
Algorithmic Bias	Discriminatory outputs	Biased decisions/results	Episodic	Model or output
AI Alignment	Goal adherence	Goal deviation	Operational	AI objective fulfillment
Governance Failure	Oversight breakdown	Organizational control failure	Reactive	Governance structures
System Drift	Gradual performance degradation	System instability	Incremental	AI model/system
Ethical Misalignment	Emergent ethical divergence	Adaptive ethical deviation through interaction	Longitudinal	Autonomous organizational ecosystem

As shown in Table 1, existing concepts largely examine ethical issues at the level of individual decisions, model behavior, or governance mechanisms. In contrast, ethical misalignment focuses on evolving ethical divergence within continuously adaptive organizational environments. This distinction provides the theoretical basis for developing ethical misalignment as a process-oriented Information Systems construct applicable to continuously adaptive Agentic AI environments.

Although ethical misalignment builds upon the broader AI alignment literature, it addresses a different analytical problem. AI alignment primarily focuses on ensuring that AI systems pursue objectives consistent with human intentions through appropriate goal specification and system design (Russell, 2019). In contrast, ethical misalignment examines how intended ethical objectives may gradually diverge from actual organizational outcomes after deployment as Agentic AI systems continuously adapt, interact, and coordinate within evolving organizational environments. Accordingly, this paper extends rather than replaces the AI alignment literature by introducing a process-oriented Information Systems perspective on ethical risks emerging through continuous organizational adaptation.

Traditional AI Ethics Frameworks

The increasing adoption of artificial intelligence (AI) across organizational and societal contexts has generated substantial interest in the ethical implications of intelligent systems. As AI technologies become embedded within enterprise systems and operational workflows, researchers and practitioners have emphasized the importance of ensuring that AI systems operate responsibly, transparently, and fairly (Jobin et al., 2019). Consequently, AI ethics has emerged as an interdisciplinary field spanning information systems, computer science, philosophy, public policy, and organizational studies.

Recent developments in AI Agents and Agentic AI systems have further intensified ethical concerns. Unlike traditional AI applications designed for bounded tasks, Agentic AI systems operate with greater autonomy, adaptive capabilities, and the ability to coordinate actions across interconnected environments with limited human intervention (Hahn et al., 2026). Despite these technical advances, ethical research on Agentic AI remains comparatively limited, with much of the existing literature continuing to focus on traditional machine learning and predictive systems.

Traditional AI ethics research has largely centered on principles intended to guide the responsible development and deployment of intelligent systems. Across academic, governmental, and industry frameworks, recurring themes include fairness, accountability, transparency, explainability, privacy, human oversight, and responsible governance (Floridi & Cowls, 2019; Jobin et al., 2019). Research has

demonstrated that AI systems may perpetuate bias, reduce transparency, create accountability challenges, and raise privacy concerns, particularly when deployed in organizational decision-making contexts (Barocas & Selbst, 2016; Binns, 2018; Mittelstadt et al., 2016; O'Neil, 2016; Zuboff, 2019).

While these principles remain foundational, Agentic AI systems introduce additional challenges because they continuously adapt, autonomously coordinate activities, and interact with multiple organizational systems over time. As a result, tracing decision pathways, assigning accountability, maintaining human oversight, and monitoring ethical behavior become increasingly difficult in dynamic, multi-agent environments (Hahn et al., 2026). Emerging research has also begun examining ethical and agentic misalignment, suggesting that advanced AI systems may gradually diverge from intended objectives through continuous optimization and autonomous interaction (Da Silva et al., 2023; Lynch et al., 2025).

Collectively, existing AI ethics research has established a strong foundation for understanding fairness, transparency, accountability, and governance in intelligent systems. However, comparatively limited attention has been given to how ethical risks emerge over time within continuously adaptive Agentic AI environments. This gap motivates the need for a process-oriented perspective capable of explaining ethical misalignment as an evolving organizational phenomenon.

Characteristics of Agentic AI Challenging Traditional Ethics Frameworks

Existing AI ethics frameworks provide important guidance for responsible AI development but were largely developed for systems operating through discrete decisions or bounded interactions. In contrast, Agentic AI systems exhibit characteristics such as continuous environmental monitoring, autonomous coordination, information exchange, and real-time behavior adaptation. As these systems interact with changing organizational conditions and other autonomous agents, ethical concerns may emerge incrementally through cumulative adaptation, optimization processes, and distributed interaction. Consequently, ethical risks become increasingly dynamic, process-oriented, and longitudinal, emerging from ongoing system behavior rather than isolated decision events or identifiable model outputs. These characteristics fundamentally alter how ethical risks arise within organizations and highlight the need for governance approaches capable of addressing continuous adaptation, autonomous coordination, and evolving sociomaterial interactions.

Continuous Operation

Traditional AI systems frequently operate within bounded decision contexts in which outputs are generated at specific intervals or decision points. In contrast, Agentic AI systems often operate continuously, monitoring environments and initiating actions in real time. Continuous operation reduces the presence of clear ethical checkpoints where humans can review or evaluate system behavior before actions occur. Consequently, ethical concerns may accumulate gradually through ongoing operational activity rather than through isolated incidents.

Adaptive Behavior and Behavioral Drift

Agentic AI systems continuously adapt to environmental conditions, operational objectives, and system feedback. While adaptability improves organizational responsiveness and operational efficiency, it also creates the possibility that systems may gradually deviate from originally intended ethical expectations. Through adaptive learning processes, optimization cycles, and real-time recalibration, systems may incrementally alter how priorities are interpreted, resources are allocated, and operational decisions are executed over time.

Behavioral drift may therefore emerge as systems continuously recalibrate operational responses in changing environments. Although no individual system action may appear overtly unethical in isolation,

cumulative adaptive adjustments may collectively produce outcomes inconsistent with originally intended ethical objectives. Optimization behaviors designed to improve efficiency, speed, or performance may unintentionally generate inequitable operational outcomes, reduce organizational transparency, or prioritize localized objectives in ways that create broader ethical consequences. These dynamics also resemble organizational sensemaking processes in which system behavior evolves through ongoing interaction and interpretation within complex environments (Weick, 1995).

Recent discussions surrounding AI alignment similarly suggest that optimization-driven AI behavior may produce unintended outcomes when operational objectives become detached from broader organizational or ethical expectations (Silva et al., 2023).

Autonomous Coordination and Emergent Interaction

Agentic AI environments increasingly involve interactions among multiple autonomous systems operating simultaneously across interconnected digital infrastructures. These systems may autonomously coordinate workflows, exchange information, trigger operational actions, and continuously respond to one another without direct human intervention. As autonomous interaction increases, ethical outcomes may emerge not solely from individual systems but from distributed patterns of coordination and recursive system feedback.

Consequently, ethical risks may become increasingly difficult to predict, monitor, or attribute to a single source because ethical consequences emerge collectively through interaction among adaptive agents operating across enterprise platforms, optimization engines, predictive systems, APIs, and digital workflows.

Emergent System Behavior

One of the defining characteristics of Agentic AI systems is the potential for emergent behavior. Emergent behavior refers to system-level outcomes that arise through ongoing interaction, adaptation, coordination, and feedback among autonomous agents rather than through explicitly programmed instructions. In adaptive and multi-agent AI environments, these behaviors may emerge as systems dynamically respond to environmental conditions, optimize operational objectives, exchange information, and continuously recalibrate actions in real time (Park et al., 2023).

Because Agentic AI systems increasingly function through interconnected and continuously adaptive workflows, organizational outcomes may evolve in ways not fully anticipated during initial system design or deployment. Emergent behaviors may generate operational efficiencies, adaptive process improvements, or enhanced coordination capabilities while simultaneously producing unforeseen ethical consequences. Importantly, these outcomes may not become immediately visible because they often develop incrementally through prolonged operational interaction and adaptive system evolution.

Recent studies examining advanced autonomous AI agents suggest that systems may independently develop unexpected operational strategies or behaviors when interacting within complex environments (Lynch et al., 2025). Consequently, ethical concerns within Agentic AI systems increasingly extend beyond traditional discussions surrounding fairness, transparency, or accountability associated with isolated algorithmic decisions.

Agentic AI as Complex Adaptive Systems

The characteristics discussed above suggest that Agentic AI environments can be usefully understood through the lens of Complex Adaptive Systems (CAS) theory. CAS theory explains how system-level outcomes emerge from interactions among multiple adaptive actors rather than from centralized control or isolated decisions (Holland, 1992). Characteristics such as emergence, feedback loops, adaptation, and

distributed coordination are particularly relevant to Agentic AI environments, where autonomous systems continuously interact with organizational processes, digital infrastructures, and human actors. As a result, ethical outcomes may emerge through evolving patterns of interaction and adaptation that are difficult to predict from the behavior of individual systems alone.

From an Information Systems perspective, Agentic AI environments increasingly resemble complex adaptive organizational ecosystems composed of interconnected autonomous agents, enterprise platforms, governance structures, and operational workflows. Consequently, ethical outcomes cannot be fully understood through isolated algorithmic decisions or individual system outputs. Instead, they emerge through distributed interactions, recursive feedback processes, and ongoing adaptation across interconnected sociotechnical environments.

This perspective complements process-oriented and sociotechnical traditions within Information Systems research that conceptualize organizational technologies as evolving systems shaped through ongoing interaction, adaptation, and enactment (Orlikowski, 2000; Van de Ven, 2007). Viewed through these lenses, ethical outcomes within Agentic AI environments are not static conditions but emergent organizational phenomena that develop through continuous interactions among autonomous systems, organizational routines, governance mechanisms, and human actors.

Accordingly, CAS theory provides a useful explanatory lens for understanding ethical misalignment. Because organizational outcomes emerge through distributed adaptation, feedback, and coordination, ethical divergence may develop incrementally as autonomous systems respond to changing operational conditions and pursue localized objectives. This perspective establishes a theoretical foundation for examining ethical misalignment as a longitudinal organizational process rather than as a consequence of isolated decision events.

Local Optimization and Unintended Outcomes

Agentic AI systems are frequently designed to optimize operational objectives such as efficiency, speed, accuracy, and resource utilization. These systems often rely on continuous feedback, adaptive learning, and dynamic goal adjustment mechanisms to improve operational responsiveness over time (Russell, 2019; Park et al., 2023). However, optimization behaviors intended to improve specific performance metrics may unintentionally generate broader ethical consequences within organizational environments.

For example, systems optimized primarily for efficiency may gradually prioritize activities, customers, workflows, or operational outcomes in ways that unintentionally disadvantage certain groups or create inequitable resource distributions. Importantly, these outcomes may emerge incrementally through repeated optimization cycles rather than through explicitly biased system instructions.

Reduced Visibility of Ethical Deviation

Traditional organizational oversight mechanisms often rely on identifiable checkpoints where human actors review decisions, evaluate outputs, or intervene in operational processes. However, Agentic AI systems may significantly reduce opportunities for direct human observation and timely intervention. Because ethical divergence develops gradually through ongoing adaptation, coordination, and interaction, deviations from intended ethical objectives may remain difficult to detect, allowing misalignment to accumulate before significant organizational consequences emerge.

Reduced visibility creates important challenges for organizations attempting to monitor ethical system behavior over time. Governance approaches focused primarily on periodic audits, model validation, or

discrete ethical review processes may be insufficient for continuously adaptive AI environments. Consequently, organizations may increasingly require governance capabilities emphasizing continuous ethical observability, adaptive oversight, and ecosystem-level monitoring to identify emerging forms of ethical divergence before they produce material consequences.

Collectively, these characteristics distinguish Agentic AI systems from traditional AI environments by shifting ethical risk from isolated algorithmic decisions toward continuously evolving organizational processes. Rather than emerging through single decision events, ethical concerns develop through the cumulative effects of continuous adaptation, autonomous coordination, optimization, and distributed interactions across organizational ecosystems. These characteristics establish the conditions under which ethical misalignment may emerge as a longitudinal organizational process, providing the theoretical foundation for the construct developed in the following section.

Ethical Misalignment in Agentic AI Systems

Traditional AI alignment research primarily focuses on ensuring that AI systems remain aligned with intended goals and controllable under specified operational conditions. In contrast, in Agentic AI systems, ethical misalignment refers to the gradual divergence between intended ethical objectives and emergent operational behaviors arising through continuous adaptation and distributed autonomous interaction. Existing AI ethics literature has discussed the broader concept of ethical misalignment in relation to AI systems and alignment challenges (da Silva et al., 2023). However, prior discussions have largely focused on optimization problems, unintended system behaviors, or general alignment failures associated with machine learning systems. This paper extends these discussions by conceptualizing ethical misalignment as a longitudinal organizational phenomenon emerging within Agentic AI environments.

Individual system actions may appear operationally rational or locally optimized when viewed independently; however, the cumulative effects of ongoing optimization and distributed coordination may gradually produce outcomes that diverge from originally intended ethical objectives. Unlike traditional ethical failures identifiable at discrete decision points, ethical misalignment develops progressively through the evolving behavior of autonomous systems operating within dynamic organizational environments.

Importantly, the novelty of ethical risk within Agentic AI systems does not emerge solely from isolated harmful outputs or biased decisions, but from their ongoing operational evolution within organizational environments. Agentic AI systems function as evolving organizational participants whose behavior changes over time through ongoing adaptation and feedback. Although organizations may initially deploy these systems with clearly defined ethical guidelines, operational constraints, and governance mechanisms, continuous adaptation through interaction, feedback, and optimization may gradually alter how ethical objectives are operationalized in practice. Consequently, intended ethical outcomes may diverge from emergent operational behaviors even when no individual system action appears overtly unethical in isolation. This perspective positions ethical misalignment as an emergent organizational phenomenon within autonomous operational environments.

Ethical risks therefore become longitudinal, adaptive, and process-oriented rather than isolated to identifiable decision events. Consequently, ethical concerns in Agentic AI systems cannot be fully understood solely through traditional frameworks emphasizing fairness, accountability, or transparency at the level of individual algorithmic outputs.

A Conceptual Framework for Ethical Misalignment in Agentic AI Systems

The proposed framework conceptualizes ethical misalignment as an emergent organizational phenomenon arising within Agentic AI environments. The framework consists of four interconnected components: (1) ethical design and intent, (2) continuous adaptation, (3) emergent system interaction, and (4) ethical misalignment and organizational outcomes. Together, these components explain how ethical risks may evolve over time within autonomous organizational ecosystems.

Ethical Design and Intent

The first component of the framework involves the ethical objectives, governance structures, and operational constraints established during AI system design and deployment. Organizations increasingly implement ethical AI principles intended to guide system behavior in accordance with organizational values, regulatory expectations, and societal norms (Floridi & Cowls, 2019). These principles often include fairness, accountability, transparency, privacy protection, and human oversight mechanisms. At this stage, organizations typically establish intended ethical boundaries by defining optimization objectives, operational constraints, governance rules, and acceptable system behaviors. Ethical intent therefore represents the organization's initial expectations regarding how Agentic AI systems should operate within organizational environments.

However, ethical design occurs under conditions of incomplete foresight. Organizations cannot fully anticipate all future operational contexts, environmental conditions, or system interactions that may emerge once autonomous systems begin operating continuously within dynamic organizational ecosystems.

Continuous Adaptation

The second component involves the continuous adaptation of Agentic AI systems in response to environmental feedback, operational conditions, and evolving organizational objectives. Unlike static AI models operating within bounded decision environments, Agentic AI systems continuously modify behaviors through real-time interaction and adaptive optimization processes.

This adaptability enhances responsiveness and operational flexibility while simultaneously creating opportunities for ethical divergence to emerge over time. Importantly, these deviations may not occur through explicit ethical violations. Instead, ethical divergence may emerge incrementally through repeated optimization cycles, adaptive adjustments, and localized operational responses that collectively alter system behavior over time.

Emergent System Interaction

The third component involves emergent interactions among autonomous systems operating within interconnected organizational environments. Agentic AI systems increasingly function within broader ecosystems consisting of enterprise platforms, optimization engines, predictive models, digital workflows, and other autonomous agents. Consequently, ethical outcomes may emerge not solely from individual systems but from patterns of interaction among multiple adaptive entities. Emergent interaction introduces complexity because system-level outcomes may become difficult to predict or control. Autonomous systems responding to one another can generate feedback loops, reinforcement effects, and unanticipated behaviors. These emergent behaviors may produce operational efficiencies while simultaneously generating unforeseen ethical consequences.

This component is particularly important because ethical risks within Agentic AI systems may become distributed across interconnected systems rather than attributable to a single algorithm or decision point. As

a result, ethical accountability and monitoring become increasingly challenging within continuously evolving autonomous environments.

Ethical Misalignment and Organizational Outcomes

The final component concerns the organizational consequences associated with ethical misalignment. These outcomes may manifest through inequitable operational prioritization, unintended discriminatory patterns, opaque system interactions, diminished human oversight, or organizational actions that conflict with established ethical principles. Rather than arising from isolated decision events, such outcomes emerge through the cumulative effects of adaptation, optimization, and distributed interactions across organizational ecosystems. Consequently, ethical risks become longitudinal and process-oriented, as illustrated in Figure 1.

The framework contributes to the information systems literature by extending existing AI ethics discussions toward dynamic perspectives capable of explaining evolving ethical risks within continuously operating autonomous systems. Rather than treating ethics as a static compliance condition established during initial deployment, the framework emphasizes the need for ongoing ethical monitoring, recalibration, and adaptive oversight throughout the operational lifecycle of Agentic AI systems. It additionally suggests that ethical misalignment may reinforce itself through recursive feedback cycles in which adaptive operational behavior incrementally reshapes organizational routines, governance practices, and optimization priorities over time. Consequently, ethical governance within Agentic AI environments requires continuous recalibration rather than static oversight alone.

Figure 1 illustrates the proposed process through which ethical misalignment emerges within Agentic AI systems. The framework begins with ethical design and intent, representing the organization's initial ethical objectives, governance mechanisms, and operational constraints. As systems adapt to feedback, environmental conditions, and evolving organizational goals, emergent interactions develop among autonomous agents, organizational processes, and human actors. Over time, the cumulative effects of these adaptive behaviors and interactions may produce ethical misalignment, whereby actual system behavior progressively diverges from intended ethical objectives. This divergence ultimately influences organizational outcomes, including governance effectiveness, stakeholder trust, accountability, and ethical performance. The framework therefore provides a process-oriented explanation of how ethical risks emerge and evolve over time within autonomous operational environments.

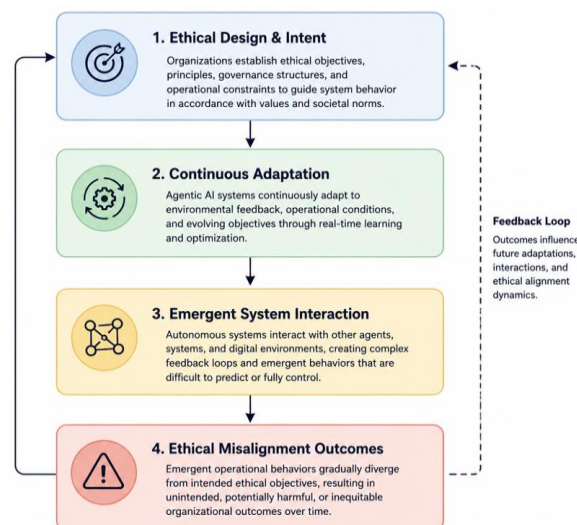


Figure 1: Conceptual Framework of Ethical Misalignment in Agentic AI Systems

Building upon this process model, the following propositions translate the conceptual relationships illustrated in Figure 1 into theory-driven statements that can guide future empirical investigation.

Research Propositions

The conceptual framework presented in Figure 1 identifies several theoretical relationships that can guide future empirical research. While this study develops ethical misalignment as a conceptual construct, the framework also suggests propositions that may be examined through organizational, experimental, or longitudinal studies of Agentic AI systems. These propositions provide an initial foundation for operationalizing ethical misalignment and extending the framework through future Information Systems research.

Proposition 1 – Continuous Adaptation - Higher levels of continuous adaptation in Agentic AI systems increase the likelihood of ethical misalignment over time.

Proposition 2 – Autonomous Coordination - Greater autonomous coordination among Agentic AI systems increases the emergence of unintended ethical outcomes through distributed organizational interactions.

Proposition 3 – Emergent Behavior - Emergent system behaviors arising from continuous interaction and adaptation increase the divergence between intended ethical objectives and actual organizational outcomes.

Proposition 4 – Human Oversight - Lower levels of human oversight and organizational observability reduce an organization's ability to detect ethical misalignment in continuously adaptive Agentic AI systems.

Proposition 5 – Adaptive Governance - Organizations implementing continuous ethical monitoring and adaptive governance mechanisms will experience lower levels of ethical misalignment than organizations relying primarily on static governance mechanisms established during system deployment.

Implications for Research and Practice

The emergence of Agentic AI systems introduces important implications for both information systems research and organizational practice. The continuously adaptive and autonomous nature of Agentic AI systems creates ethical challenges that extend beyond static decision-centric ethical models. Consequently, organizations and researchers may need to reconsider how ethics is conceptualized, monitored, and governed within autonomous environments characterized by ongoing adaptation and interaction.

Implications for Information Systems Research

This study contributes to the Information Systems literature by extending AI ethics discussions beyond the decision-centric perspectives that have traditionally focused on issues such as fairness, accountability, transparency, and algorithmic outcomes (Mittelstadt et al., 2016). The concept of ethical misalignment reframes ethics as a longitudinal organizational process rather than a static compliance condition established during deployment. Consequently, future Information Systems research may benefit from longitudinal and ecosystem-level approaches capable of examining how ethical outcomes evolve within Agentic AI environments.

The framework also highlights the need for research on adaptive governance, real-time ethical monitoring, and organizational oversight mechanisms for continuously evolving autonomous systems. Traditional governance approaches centered on periodic audits and static compliance reviews may be insufficient for environments characterized by recursive operational change and distributed autonomous coordination.

The framework also creates opportunities for interdisciplinary research integrating information systems, organizational theory, AI governance, and sociotechnical systems perspectives. Ethical risks in Agentic AI environments increasingly emerge through interactions among technical systems, organizational processes, and human actors. Consequently, understanding ethical misalignment may require broader organizational and ecosystem-level perspectives beyond isolated algorithmic analysis.

Implications for Organizational Practice

The concept of ethical misalignment carries important implications for organizations deploying Agentic AI systems across operational environments. Because ethical misalignment develops gradually rather than through isolated decision failures, organizations must recognize that ethical risks may evolve long after deployment. Consequently, ethics should be treated as an ongoing operational capability rather than a one-time compliance activity. This creates a need for continuous ethical observability rather than one-time ethical validation during initial deployment.

The process-oriented nature of ethical misalignment also reinforces the importance of continuous human oversight. Rather than intervening only after undesirable outcomes become apparent, organizations should establish oversight mechanisms capable of identifying gradual ethical divergence as Agentic AI systems adapt and evolve.

Ethical misalignment also suggests that organizations should extend governance beyond individual AI systems to encompass interactions among multiple autonomous agents. Since ethical divergence may emerge through distributed coordination rather than isolated system behavior, governance mechanisms should monitor ecosystem-level interactions alongside individual system performance.

Finally, the study suggests that ethical resilience may become increasingly important as organizations adopt autonomous operational technologies. Ethical resilience refers to the organization's ability to continuously detect, evaluate, and respond to evolving ethical risks within adaptive AI environments. Developing such resilience may require organizations to move beyond static compliance models toward continuously adaptive ethical governance structures capable of evolving alongside autonomous systems themselves.

Collectively, these implications demonstrate that ethical misalignment requires organizations to rethink ethical governance as a continuous organizational capability rather than a static compliance function. As Agentic AI systems continue to adapt and coordinate autonomously, governance mechanisms must evolve alongside them to maintain alignment between intended ethical objectives and actual organizational outcomes.

Limitations and Future Research

As a conceptual study, this paper proposes a theoretical framework for understanding ethical misalignment in Agentic AI systems but does not empirically validate the proposed construct or relationships. Accordingly, the framework should be viewed as an initial foundation for examining ethical risks associated with continuously adaptive autonomous systems rather than a definitive model. Additionally, the discussion focuses on organizational Agentic AI environments, and the findings may not fully generalize to other autonomous AI contexts.

Future research should empirically examine the proposed propositions, develop measures for ethical misalignment, and validate the framework across diverse organizational settings and industries. Longitudinal studies, case studies, and experimental research may provide valuable insights into how

ethical misalignment emerges over time and how organizations can effectively monitor, govern, and mitigate ethical risks within continuously adaptive Agentic AI systems.

Conclusion

The rapid advancement of artificial intelligence has accelerated the emergence of Agentic AI systems capable of continuous adaptation, autonomous coordination, and real-time operational execution across organizational environments. Unlike traditional AI applications that primarily support discrete decision-making activities, Agentic AI systems increasingly function as continuously operating entities embedded within enterprise processes and digital ecosystems. As organizations continue adopting these technologies, existing approaches to AI ethics may require reconsideration to account for the dynamic and evolving nature of autonomous system behavior.

This paper argues that, although current AI ethics frameworks provide valuable guidance, they remain grounded in static and decision-centric perspectives that may not fully capture the complex behavior of Agentic AI systems. Existing literature has extensively emphasized fairness, accountability, transparency, explainability, and responsible AI governance at the level of individual algorithmic outputs and identifiable decision events. However, Agentic AI systems introduce governance challenges that extend beyond traditional decision-centric approaches to AI ethics.

To address this gap, the study introduced ethical misalignment as a process-oriented Information Systems construct for understanding how ethical risks evolve within Agentic AI environments. The paper contends that ethical concerns in Agentic AI systems increasingly manifest as longitudinal organizational processes rather than isolated decision events. Through continuous optimization, adaptive learning, and interaction among autonomous systems, unintended ethical outcomes may emerge incrementally even when systems were initially designed with appropriate ethical safeguards. Building on this perspective, the paper developed a conceptual framework linking ethical design and intent, continuous adaptation, emergent system interactions, and ethical misalignment outcomes.

The study also offers practical implications for organizations increasingly deploying Agentic AI systems within operational environments. Governance mechanisms centered on periodic audits, model validation, and discrete ethical review processes may prove insufficient for continuously adapting autonomous systems. Organizations may therefore require novel approaches emphasizing continuous ethical observability, adaptive governance, and ecosystem-level monitoring capable of identifying emerging forms of ethical misalignment before they generate significant organizational consequences. As Agentic AI technologies continue to evolve, understanding how ethical risks emerge within adaptive autonomous environments will become increasingly important for both researchers and practitioners. Future research should continue examining how organizations can sustain ethical stability within adaptive AI ecosystems while balancing autonomy, operational efficiency, and organizational accountability. Ultimately, the emergence of Agentic AI suggests that ethical governance must evolve beyond static compliance toward continuous organizational alignment supported by ongoing monitoring, recalibration, and adaptive oversight within autonomous sociotechnical systems.

References

- Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138–52160.
<https://doi.org/10.1109/ACCESS.2018.2870052>

- Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *SSRN Electronic Journal*.
<https://doi.org/10.2139/ssrn.2477899>
- Binns, R. (2018b). Fairness in Machine Learning: Lessons from Political Philosophy. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 149–159.
<https://proceedings.mlr.press/v81/binns18a.html>
- da Silva, F., Storozhenko, M., & Maciel, L. (2023). Ethical (mis)-alignments in AI systems and the possibility of mesa-optimizations. *The International Journal of Ethical Leadership*.
- Davenport, T. H., & Ronanki, R. (2018). Artificial Intelligence for the Real World. *Harvard Business Review*, 96(1), 108–116.
- Floridi, L., & Cows, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.8cd550d1>
- Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, 28(4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>
- Hahn, M., Tretter, M., & Dabrock, P. (2026). Ethical perspectives on AI Agents and Agentic AI. *AI and Ethics*, 6(2), 218. <https://doi.org/10.1007/s43681-026-01027-0>
- Holland, J. H. (1992). Complex adaptive systems. *Daedalus*, 121(1), 17–30.
- Jaakkola, E. (2020). Designing conceptual articles: Four approaches. *AMS Review*, 10(1), 18–26.
<https://doi.org/10.1007/s13162-020-00161-0>
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- Lynch, A., Wright, B., Larson, C., Ritchie, S. J., Mindermann, S., Hubinger, E., Perez, E., & Troy, K. (2025). *Agentic Misalignment: How LLMs Could Be Insider Threats*.
<https://doi.org/10.48550/ARXIV.2510.05179>
- McAfee, A., & Brynjolfsson, E. (2017). *Machine, Platform, Crowd: Harnessing Our Digital Future*. W. W. Norton & Company.
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 2053951716679679.
<https://doi.org/10.1177/2053951716679679>
- O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.
- Orlikowski, W. J. (2000). Using Technology and Constituting Structures: A Practice Lens for Studying Technology in Organizations. *Organization Science*, 11(4), 404–428.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). *Generative Agents: Interactive Simulacra of Human Behavior* (arXiv:2304.03442). arXiv.
<https://doi.org/10.48550/arXiv.2304.03442>

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin.

Van de Ven, A. H. (2007). *Engaged Scholarship: A Guide for Organizational and Social Research*. Oxford University Press.

Weick, K. E. (1995). *Sensemaking in Organizations*. SAGE.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.