Journal of International DBA Studies (JIDS)
Vol. 2 No. 001 (2026) https://eaj.ebujournals.lu/index.php/JIDS/index

Governing AI Agents: Bounded Autonomy and Human Oversight

by
Prasad Modali*

Abstract

The sudden arrival of Artificial Intelligence (AI), is shifting both scholarly and governmental focus away from issues of technological feasibility to greater concern regarding its use and oversight. AI, which promises great advancements in operational efficiency, scalability and adaptive learning, however presents many challenges as it is used in large-scale applications, including those of ethics and compliance with law; accountability and the ability of humans to exercise meaningful control. Traditional decision-support systems differ from AI agents because they are capable of independently initiating action, coordinating workflow processes, and modifying behavior based upon past experience. These characteristics create new obstacles for existing governance and oversight mechanisms. Research continues to emphasize technical performance and model-based protections against unwanted behavior by AI agents. There remain substantial gaps in governance structures that can support autonomous AI while maintaining some level of institutional accountability. This paper proposes a Bounded Autonomy Governance Framework for AI Agents to be developed using autonomy as a design variable to be governed during all phases of the development process. The proposed framework includes several key components such as calibrated levels of autonomy for AI agents, human-in-the-loop and human-on-the-loop mechanisms to include ethical guardrails, and compliance-by-design provisions throughout the entire life cycle of an AI agent. Using a multi-disciplinary body of literature relating to AI governance, digital infrastructure, and organizational systems, the author views AI agents as social and technical actors embedded in institutions. The paper also builds from the bounded rationality and absorptive capacity literature to strengthen the theoretical case for studying bounded autonomy. Through both theoretical and practical design considerations —including an applied case walkthrough—this paper provides actionable recommendations for policymakers, platform designers, and industry practitioners wishing to responsibly develop and implement AI agents at scale.

Keywords: AI agents; bounded autonomy; AI governance; human oversight; responsible AI; socio-technical systems; compliance-by-design; digital infrastructure

* Prasad A. Modali is a strategy and analytics professional in India with over three decades of experience across large conglomerates, working at the intersection of business analysis, digital transformation, and enterprise strategy. His research interests focus on governance of emerging technologies, responsible deployment of AI systems, and the institutional frameworks required to align digital innovation with organizational and societal objectives.

Introduction

In recent years, AI has evolved significantly from the use of experimental computational techniques to utilize systems in a number of areas including organizational operations, economic activities and providing public services. Advancements made in AI technology have also led to the creation of AI agents—systems that can perform tasks based on not only analytical and recommendatory information, but undertake autonomous action in multiple parts of complex workflows (Russell & Norvig, 2021; Wooldridge, 2021).

As such, the current state of AI represents a qualitative difference compared to past forms of AI, leading to new governance issues that previous frameworks were unable to handle. Past generations of AI operated as decision support tools. They provided recommendations that human decision makers would then review and take appropriate action based on those recommendations. In doing so, it was clear that responsibility rested solely with the humans involved. Unlike these past forms of AI, AI agents do have the authority to take independent actions, orchestrate workflows and make significant decisions with little to no direct human involvement.

Some examples of agentic systems include customer service systems that can respond to customer inquiries and solve problems without further input from customers or other human employees; workflow-orchestrating systems that can manage business processes; and infrastructure systems that can automatically distribute network or energy resources in real-time.

The rising importance of agentic systems are reflected in the increase in investments by companies across all industries. Companies are testing agentic systems as a means to increase efficiency, scalability and responsiveness. However, governments and regulatory bodies are expressing concerns in the delegation of decision making authority to machines. These concerns are especially evident in high-risk areas such as finance, healthcare, public administration and critical infrastructure where autonomous actions taken by agentic systems could negatively impact rights, safety and access to essential services.

While agentic systems hold great potential for improving efficiency and reducing costs for organizations, there are risks associated with their deployment that cannot be mitigated using conventional governance structures. Since agentic systems operate continuously, adapt to changing circumstances, and learn from interactions with users and data sources, their behavior will likely evolve over time in unpredictable ways.

Additionally, errors or bias that occur during individual decisions made by agentic systems can compound over time resulting in unforeseen consequences. Finally, because agentic systems can operate autonomously, determining responsibility for any negative consequences of their actions can be problematic, potentially resulting in ambiguity as to who should bear responsibility if something goes awry.

This paper proposes bounded autonomy as a governance concept that permits effective operation of agents under well-defined technical, ethical and institutional

boundaries treating autonomy as a variable that can be adjusted, constrained and evaluated continually.

Motivation/Problem Definition

The rapid dissemination of AI agents has out-paced the development of adequate governance mechanisms to ensure accountable exercise of their autonomy. Traditional AI deployments relied on human decision makers as a "final check point," while increasingly today's AI agents are deployed across connected systems with virtually no real-time supervision — thereby creating a governance gap between the capabilities of oversight models originally developed for decision support tools and the abilities of autonomous systems that continuously adapt and generate outputs.

Granting discretionary decision making power to machines presents risks qualitatively different than traditional process automation. By transitioning from advisory AI to agentic AI, the locus of control and responsibility within organizations (Wooldridge, 2021) shifts fundamentally necessitating governance responses that extend beyond existing technical safeguards and ethical principles.

Shortcomings of Current Governance Approaches

Existing approaches to governance for AI fall broadly into two categories. First, they emphasize principle-based ethics, defining concepts like fairness, transparency, accountability etc., which are normative and important, however they are typically vague and hard to translate into concrete constraints applicable to autonomous systems. Second, many approaches focus on technical controls such as model validation auditing and monitoring — necessary but insufficient when systems operate autonomously interacting with their environment.

In practice, organizations often establish governance structures reactively in response to either adverse events or regulatory pressures. A reactive approach is particularly hazardous for agentic systems given the possibility for gradual changes to their behaviour that may only become apparent after some length of time.

Research Objectives and Contributions

The overarching goal of this research is to create a governance structure that will allow organizations to utilize agentic systems at scale without sacrificing accountability or ethical oversight. Bounded autonomy will serve as both the conceptual framework and practical solution for addressing this objective.

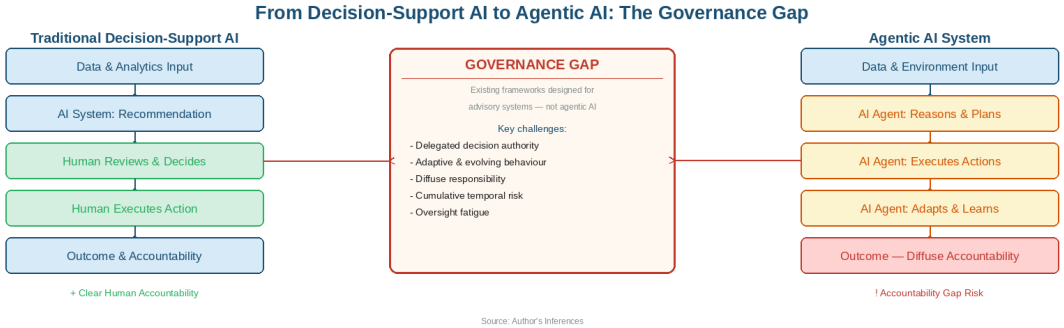
The research contributes three key findings: (1) a theoretically grounded reframing of autonomy as a governable continuum; (2) an actionable four layer governance framework; and (3) an illustrative walkthrough example describing how the proposed framework can be applied to real world situations.

Significance for Research and Practice

For researchers, practitioners, and policymakers, the bounded autonomy framework offers the possibility of empirically testing and comparatively analyzing bounded autonomy across sectors and different regulatory regimes. For practitioners and policymakers, it also represents a conceptual guide for implementing governance practices that support preserving trust, legitimacy and human agency while promoting innovative applications of AI.

Figure 1

Transition from Decision-Support AI to Agentic AI: The Governance Gap



The illustration shows the quantitative shift from existing decision-support AI systems in which humans continue to have complete control over decisions made to systems that can autonomously act throughout complex processes.

Conceptualizing AI Agents as Socio-Technical Actors

Conceptually, AI agents exist neither as discrete technical objects nor as social entities. Rather, they exist as socio-technical actors embedded within organizations, institutions and regulatory frameworks. The behavior exhibited by AI agents emerge from the interplay between the technical capabilities of their algorithms, the actions and responses of humans interacting with the system, and governance frameworks that define the parameters for acceptable behaviour (Floridi et al., 2018; Gasser & Almeida, 2017). Thus, the framing of AI agents as socio-technical actors requires a shift away from focusing solely on assessing technical performance toward examining the institutional factors shaping how AI agents behave and how they should be regulated. Viewing AI agents as integrated systems allows for examination of four primary dimensions of an AI agent:

1. Perception Layer: captures information related to either digital or physical environments;
2. Cognitive Layer: responsible for reasoning/decision-making, planning, learning, etc.;
3. Action Layer: performs tasks or interacts with other systems;
4. Feedback Mechanism: assesses outcome(s) and adjusts future behaviour accordingly. Integrating these dimensions creates emergent behaviour in AI agents functioning in real-world settings that cannot be explained solely through technical evaluations.

AI agents operate within organizational workflow structures where their actions intersect with human activities, institutional norms and regulations. An example would

include an enterprise-based AI agent focused on optimizing procurement processes that could influence supplier selection, pricing negotiation and compliance outcomes.

Similarly, a public sector-based agent used for determining welfare benefit eligibility could influence individuals' access to health care services or welfare benefits. As such, their impact are no longer measured solely through process efficiencies but also by fairness, accountability and legitimacy.

Types of AI Agents

Past research identified three types of agents based on functionality and autonomy levels; each type poses unique challenges to regulation (Wooldridge, 2021):

- Task Specific Agent: performs narrow function including but not limited to scheduling, routing and/or resolving queries.
- Workflow Orchestration Agent: coordinates multiple tasks across systems and departments.
- Decision Execution Agent: converts analytical output to action such as approvals of transactional activity or allocation of resources.
- Multi-Agent System: multiple agents communicate and coordinate resulting in collective outcomes unattributable to individual agents.

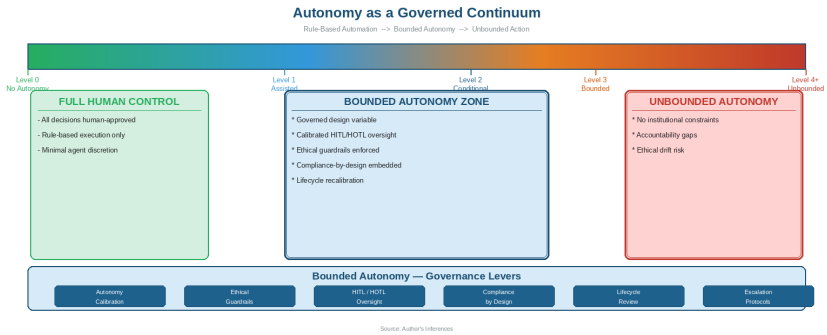
Each of the above types present unique regulatory issues. Examples include task specific agents requiring minimal supervision relative to decision execution and multi-agents raising additional complexity regarding accountability and control especially when agents produce outcomes that result from interaction with each other.

Levels of Autonomy

On one end of the spectrum are fully supervised systems that require human approval for all actions taken. On the other hand are systems that can independently execute tasks without requiring human approval within specified domain boundaries. Many current AI agents fall somewhere between. Viewing autonomy as a continuum instead of being viewed as an either/or concept allows for the development of more granular governance strategies that tailor oversight and constraints to the level of risk and potential impact (Russell & Norvig, 2021). Notably this conceptualization frames autonomy as a governance variable - one that can be formally established by regulation, monitored and revised.

Figure 2

Autonomy as a Governed Continuum: From Rule-Based Automation to Bounded Autonomous Action



Organizational Embeddedness and Institutional Context

Given that AI agents are situated within institutional frameworks there exists a distribution of responsibility for their actions across designer/deployer/oversight bodies thereby heightening the risk of accountability gaps unless clear definitions for governance mechanisms are articulated (Kroll et al., 2017; Rahwan, 2018). Framing AI agents as socio-technical actors emphasizes the necessity for assigning/responsibility for accountability rather than treating accountability as an after-the-fact consideration.

Human–Agent Interaction and Role Reconfiguration

When we deploy an AI agent, it changes what humans do inside organizations. Instead of executing tasks directly, humans will be in charge of supervising the agents that have been assigned to execute tasks, dealing with exceptions, and making strategic decisions. These role changes create cognitive and organizational problems such as fatigue from overseeing others, loss of skills due to less frequent use, and confusion about who is ultimately in charge. The more autonomous the agents become, the harder it becomes for humans to keep track of the situation, which decreases the effectives of supervision and creates distrust. Therefore, effective governance must address the issues created by the interactions between humans and AI agents, and ensure that humans continue to play meaningful roles when agents' autonomy levels increase.

Risk, Power, and Asymmetry in Agentic Systems

Autonomous AI agents can also create new types of asymmetry of power. When large numbers of organizations deploy autonomous agents they can acquire significant amounts of control over markets, information flow or public services. However, there are many people whose lives are impacted by the actions of the autonomous agents deployed by organizations.

Review of Literature and Identified Gaps in Governance

There is a substantial body of literature concerning AI agents. There are four major categories of research on AI agents:

1. Architectures of AI agents - strong on capabilities and very weak on governance.
2. Normative and prescriptive frameworks for ethical AI - important but non-operationalized for autonomous behaviour.
3. Human-in-on/with-the-loop models - well-intended but not calibrated to the level of autonomy; and
4. Governance and regulation - high-level principles with no lifecycle specificity.

Although the existing research provides important insights into AI agent architecture, ethics and responsible AI, human-AI interaction, and governance and regulations, the research has never been synthesized into an integrated approach to governance of agential systems.

Table 1

Summary of Key Literature on AI Agents, Governance, and Identified Gaps

Title	Author(s)	Year	Study Focus	Identified Gap
Artificial Intelligence: A Modern Approach	Russell & Norvig	2021	Foundational text on intelligent agents, autonomy, and decision architectures	Treats autonomy as technical capability; no institutional governance guidance
A Brief History of Artificial Intelligence	Wooldridge	2021	Historical and conceptual evolution of AI and agent-based systems	Conceptual treatment of agent autonomy; lacks deployment governance models
Artificial Intelligence Agents: Architectures and Applications	Piccialli et al.	2025	Reviews AI agent architectures and applications across domains	Emphasizes performance; governance, oversight, and accountability largely absent
Guidelines for Human–AI Interaction	Amershi et al.	2019	Design principles for human–AI interaction and oversight	Interface-level focus; does not address scalable governance for autonomous agents
The Ethics of Algorithms	Mittelstadt et al.	2016	Ethical challenges of algorithmic decision-making	Conceptual ethics focus; no operational mechanisms for autonomous action governance
AI4People — An Ethical Framework for a Good AI Society	Floridi et al.	2018	Ethical principles for societal AI governance	Normative and principle-based; does not specify autonomy constraints in practice
OECD Principles on Artificial Intelligence	OECD	2019	International policy principles for trustworthy AI	High-level guidance; limited operationalization for agents with delegated authority
Accountable Algorithms	Kroll et al.	2017	Legal and institutional accountability for algorithmic systems	Accountability mechanisms addressed; proactive autonomy constraints absent
Society-in-the-Loop	Rahwan	2018	Embedding societal values into algorithmic systems	Conceptual governance vision; lacks concrete design frameworks for agent autonomy

AI Agent Architectures and Autonomy

The development of AI agents has been driven primarily by developments in Reinforcement Learning, Planning Systems, Multi-Agent Coordination and the pursuit of autonomy as a technological goal rather than a governance issue (Wooldridge, 2021; Russell & Norvig, 2021). The advances in capabilities of agents developed within this body of work have largely abstracted away from organizational and institutional context.

Issues of accountability, the delegation of decision authority and error handling are commonly left unaddressed.

Ethical and Responsible AI Frameworks

Ethical AI frameworks provide a set of principles including fairness, transparency, and accountability. However, most of them lack specific guidelines regarding how those principles apply to autonomous systems acting over time (Floridi et al., 2018; Mittelstadt et al., 2016; OECD, 2019). Those frameworks normally consider relatively static systems. Autonomous AI agents introduce learning, adaptation, and sequential decision making and create new types of ethical risks from cumulative behavioral patterns rather than single decisions.

Human-in-the-Loop and Human-on-the-Loop Models

Human-in-the-loop (HITL) and Human-on-the-loop (HOTL) are widely accepted as a means to mitigate the risk associated with autonomous systems. Yet there is little guidance provided by the literature on how oversight should increase with increased autonomy (Doshi-Velez & Kim, 2017; Amershi et al., 2019). Oversight intensity is infrequently scaled to either autonomy levels or contextualized risk. Over-intervention can undermine both efficiency and scalability while under-intervening may allow unacceptable harm. The need for systematic calibration of oversight intensity is a significant remaining gap.

Governance and Regulatory Perspectives

Existing governance approaches often do not account for the implications of delegating action authority, adaptive behavior, and temporal accumulation of risk (Gasser & Almeida, 2017; Kroll et al., 2017). Many policy frameworks focus on ensuring accountability and human control and thus tend to be high-level and based on principles. Fewer than a handful of organizations have integrated their governance principles into their use of AI — particularly so in terms of agent-based systems in which typical compliance mechanisms are inadequate.

Research Gap Synthesis

Governance-related aspects are considered in disparate ways throughout the various literatures on autonomy, ethics, oversight, and compliance — as if they were distinct issues rather than interdependent aspects of a singular challenge. There clearly exists a lack of comprehensive frameworks integrating these concepts throughout the entire lifecycle of an AI agent. This paper attempts to fill this gap by developing bounded autonomy as an overarching governance concept connecting delegation, constraint, oversight, and accountability into a cohesive operational framework.

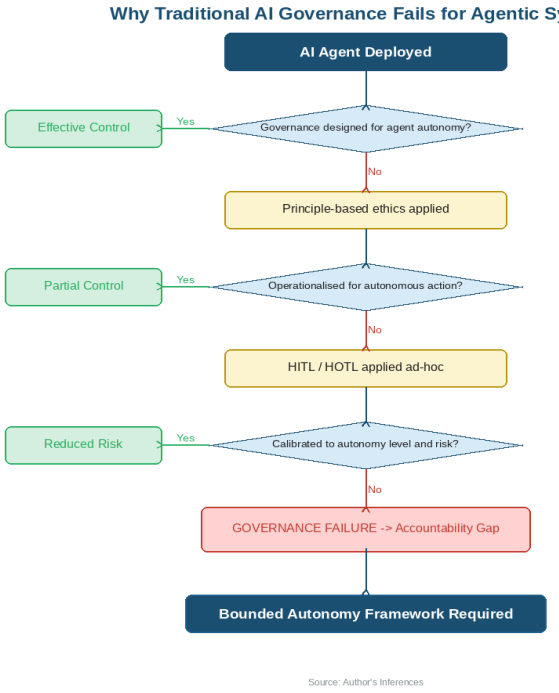
Trust, Reliance, and Automation Bias

Trust plays an important role in how humans interact with autonomous agents. Automation bias is demonstrated through the fact that humans over-rely on automated systems despite knowing that they contain errors — which is exacerbated in the

autonomous domain due to reduced verification opportunities. Conversely, insufficient trust can also result in failure to utilize an agent's full capability. Therefore, governance frameworks must consider both technical reliability and social/cognitive factors related to human-agent interaction.

Figure 3

Why Traditional AI Governance Fails for Agentic Systems: A Decision Flowchart



Research Methodology

This study uses a conceptual design science research (DSR) method. The literature synthesizing the framework was developed using a systematic process for conducting a structured bibliometric review.

Works relevant to the research objectives were identified using the above databases. Search terms used in searching were based on the keyword clusters: "AI Agents", "Bounded Autonomy", "AI Governance", "Human-In-The-Loop", "Bounded Rationality", "Absorptive Capacity", "Autonomous Systems Governance", and "Responsible AI". Papers published from 1955 – 2025 were included in the search so as to include all foundational theoretical contributions (i.e. Simon, 1955; Cohen & Levinthal, 1990), and more recent contributions related to AI governance.

Each paper had to meet inclusion criteria such that they addressed at least one of the following areas: Agent Architecture, Autonomy Design, Governance Frameworks, Human Oversight Mechanisms, Organizational Decision Theory or Regulatory Implications of AI. A total of 28 core works were reviewed and Table 2 is provided below to outline the papers within their respective theoretical domains (Gregor & Hevner, 2013; Peffers et al., 2007), but practitioners need something they can use to solve their problems. The researcher does not just develop knowledge of the problem area; he also develops a workable artifact in response to the problem, i.e. in our case the governance framework.

Design Science Research: Rigor and Validation

Table 2

Bibliometric Review: Papers Reviewed by Theoretical Domain

Author(s) & Year	Domain	Key Contribution to Framework
Simon (1955, 1957, 1972)	Bounded Rationality	Foundational constraints on decision-making; satisficing under uncertainty
March & Simon (1958)	Bounded Rationality	Organizational decision-making under cognitive and informational limits
Kahneman & Tversky (1974); Kahneman (2003)	Bounded Rationality	Heuristics and biases in human judgment; automation bias
Cohen & Levinthal (1990)	Absorptive Capacity	Foundational framework for organizational knowledge assimilation and innovation
Zahra & George (2002)	Absorptive Capacity	Re-conceptualization: potential vs. realized absorptive capacity; governance of learning
Huber (1991)	Absorptive Capacity	Organizational learning processes; knowledge acquisition and distribution
Hansen (1999)	Absorptive Capacity	Knowledge transfer across organizational units; weak-tie mechanisms
Russell & Norvig (2021)	AI Agent Architecture	Foundational agent architectures and autonomy design
Wooldridge (2021)	AI Agent Architecture	Historical evolution of AI agency; types and autonomy levels
Floridi et al. (2018)	AI Ethics & Governance	Ethical principles for AI society; basis for guardrails layer
OECD (2019)	AI Ethics & Governance	International AI governance principles; trustworthy AI policy
Kroll et al. (2017)	AI Ethics & Governance	Algorithmic accountability; compliance layer grounding
Rahwan (2018)	AI Ethics & Governance	Society-in-the-loop; embedding societal values in autonomous systems
Amershi et al. (2019)	Human-AI Interaction	Design guidelines for human-AI oversight; HITL/HOTL calibration
Doshi-Velez & Kim (2017)	Human-AI Interaction	Interpretable machine learning; explainability for accountability
Gregor & Hevner (2013); Hevner et al. (2004); Peffers et al. (2007)	Design Science Research	Methodological grounding for artefact-based framework development
Mittelstadt et al. (2016)	AI Ethics & Governance	Ethical challenges of algorithmic decision-making
Gasser & Almeida (2017)	AI Ethics & Governance	Layered AI governance model; institutional embeddedness
Piccialli et al. (2025)	AI Agent Architecture	Contemporary agent architectures and cross-domain applications

Table 2 presents the core papers reviewed through the bibliometric process. The two foundational theoretical streams—bounded rationality and absorptive capacity—are included as required scholarly grounding alongside AI governance and agent architecture literature. The purposefully broadened scope of the bibliometric study covers two theoretical fields (bounded rationality and absorptive capacity) that have not traditionally been cited in AI governance research; however, they represent key elements of the conceptual architecture of this paper. They were intentionally included in order to be structural necessities.

Bounded rationality was first described by Herbert A. Simon (1955, 1957, 1972) and further elaborated upon by March and Simon (1958) and Kahneman (2003). The theory posits that all decision making processes (human or artificial intelligence) are limited by insufficient information, limited processing ability and time constraints. Thus, autonomous AI decision making can never be completely optimized; rather, it will result in satisfactory responses within predetermined operational boundaries. Consequently, any governance structure that treats AI's ability to act independently of human intervention as unlimited or self-correcting mischaracterizes how these systems actually make decisions. All three layers of the Bounded Autonomy Framework (from autonomy thresholds in Layer 1 through to escalation protocols in Layer 3) have been structurally developed with this conceptual foundation.

Absorptive capacity was initially established by Cohen & Levinthal (1990) and has since been expanded upon by Zahra & George (2002) and Huber (1991). In terms of organizations (and correspondingly AI systems), absorptive capacity represents an organization's (AI system) ability to recognize, assimilate and utilize new external knowledge. With regard to governed AI, this implies that agents must adapt without drifting into unstable/unpredictable behaviour. This concept directly influenced the lifecycle governance overlay described in Section 5.8, where recalibration loops were specifically designed into the framework design.

More critically than previously mentioned both of the above frameworks apply equally to regulators and institutions responsible for overseeing AI as well as to the AI systems themselves. Effective governance requires acknowledgment that policymakers also operate under Bounded Rationality, therefore need institutional Absorptive Capacity to continuously integrate emerging risks and technological developments.

Crucially, both literatures are extended here beyond their original organizational contexts to introduce and justify the need to study bounded autonomy as a governance construct — one that is structurally dependent on recognising the satisficing nature of AI decision-making and the necessity of governed, selective learning as agents operate across dynamic institutional environments. The development of this framework conforms to widely accepted DSR principles whereby the principal contribution is a justified artefact (Gregor & Hevner, 2013).

Unit of Analysis and Analytical Lenses

The unit of analysis here is the AI agent viewed as a socio-technical system. Thus, while each AI agent may be viewed as an independent algorithmic or modelling entity, we view them collectively as part of organizational workflow systems that operate under ethical norms, regulatory constraints and human oversight. As such we are enabled to examine the interaction effects that occur between people, institutions and technology. Framework development was conducted using three complementary analytical lenses as presented in Table 2

Table 3

Analytical Lenses Guiding Framework Development

Analytical Lens	Purpose
Autonomy Lens	Examines degrees of decision-making authority delegated to AI agents and the conditions under which such delegation is appropriate
Governance Lens	Focuses on accountability, compliance, and auditability across the agent lifecycle
Human Oversight Lens	Evaluates mechanisms for human control and intervention, calibrated to autonomy level and contextual risk

These lenses operate concurrently to ensure that governance constructs address technical, institutional, and human dimensions simultaneously.

Governance Framework for AI Agents

The previous chapters illustrated how all prior forms of AI governance fail to provide an effective means for governing systems that will operate independently over long periods of time. As a direct response to this limitation, the Bounded Autonomy Governance Model developed herein views autonomy as the key component of governance.

Instead of merely regulating results after they have been produced, Bounded Autonomy framework provides a model for establishing boundaries regarding what autonomous agents can do when, at what level of human oversight, and through a process of ongoing evaluation—making autonomy both conditional and reversible.

Conceptual Foundations of Bounded Autonomy

Bounded Autonomy recognizes that AI agents must function with some degree of freedom from human control to create value, but also insists that such freedom is always constrained within the bounds defined by human objectives and norms. Because unlike traditional automation, AI agents use their own judgment to make context-specific decisions under uncertainty over time, governance models based on pre-defined rules alone, or post-implementation audits alone are insufficient.

A model that combines *ex ante* constraints (what agents are allowed to do), real-time control mechanisms (under what conditions humans may intervene), and ex post accountability (who is accountable for actions taken) is necessary. In addition to providing an integrated model that incorporates these different elements into a cohesive

framework, Bounded Autonomy has also established the first comprehensive framework for implementing a model of AI governance.

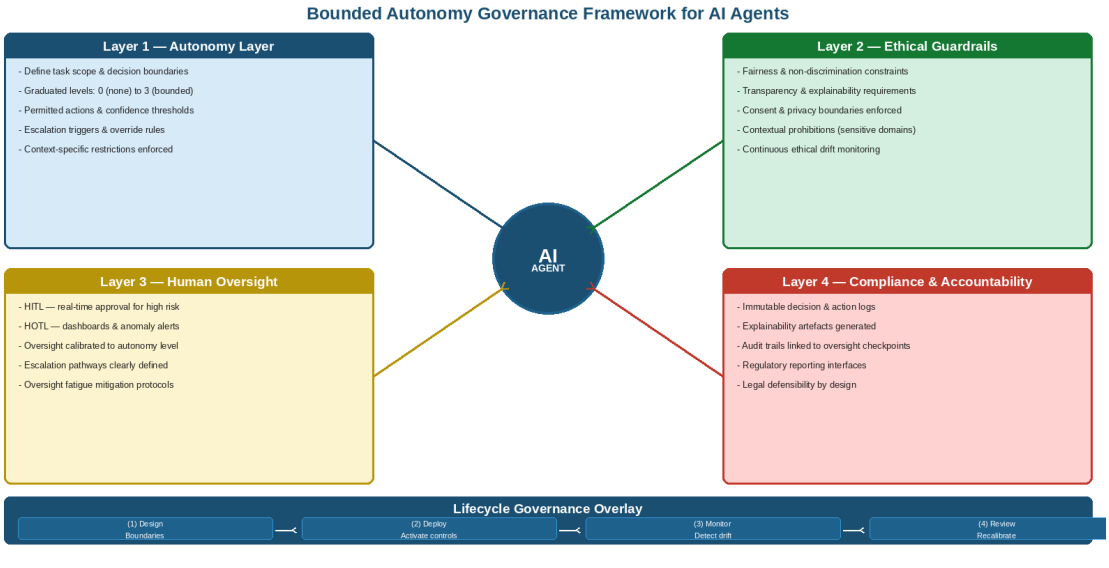
Second, the lifecycle governance framework also pulls in absorptive capacity theory as presented by Cohen and Levinthal (1990), where the absorptive capacity of an organization is its ability to 'recognize the value of new information, assimilate it, and apply it to commercial ends'. For AI systems in dynamic environments, this points to a need for governed learning—the agent must update based on new data in a selective, traceable way that builds on previously validated knowledge. Zahra and George's (2002) re-conceptualisation of absorptive capacity into its potential and realized dimensions is particularly relevant here—the distinction between the agent's capacity to acquire new knowledge and to deploy that knowledge appropriately resonates with the governance challenge of the learning-enabled AI agent "maturing" in a controlled, auditable way instead of in opaque or unstable ways. Huber's (1991) basic framework on organizational learning also lends support to the discussion in the context of designing feedback loops into the lifecycle governance overlay (Section 5.8) in which outcome data is fed back into framework recalibration. Together, bounded rationality and absorptive capacity tell us that not only AI systems can learn, but the regulatory institutions behind them are also subject to cognitive or informational constraints, the effective governance of which must build in mechanisms for iterative learning, oversight adjustment and institutional recalibration.

Framework Overview

A total of four interconnected levels comprise the Bounded Autonomy Governance model; each level exists simultaneously in operation. All four levels must remain in balance: if one level is altered (e.g., expanded agent autonomy)—the other three levels must be modified correspondingly (e.g., increased human oversight, stricter ethical constraints). The complete architecture is presented in Figure 4.

Figure 4

Bounded Autonomy Governance Framework for AI Agents: Four-Layer Architecture



Source: Author's Inferences

A "four-layer" lifecycle-governance component overlays the four layers to enable ongoing assessments of an agent's behaviour. All four layers influence and are influenced by all other layers (bi-directional arrow).

Layer 1: Autonomy Layer

The autonomy layer determines how much of a role an AI agent can take in performing tasks. It does so based upon restrictions such as: what tasks the AI agent is allowed to perform, under what conditions, and what level of independent judgment the agent may exercise. As opposed to allowing an AI agent unfettered access to every possible action, the autonomy layer outlines specific parameters for the agent including: threshold for making decisions, specific actions that can be taken by the agent, and procedures for escalating problems. The overall objective of the autonomy layer is to define how an agent will behave.

An example of how the levels of autonomy may be defined is outlined in Table 3. The majority of current implementations of AI agents fall into Levels 1 and 2; however, the framework allows for a safe transition to Level 3 if the developer demonstrates that they have successfully implemented reliable mechanisms for ensuring the appropriate use of the technology, demonstrates sufficient levels of organizational governance, and ensures that the application of the technology is appropriate in light of the context within which it is being used.

Table 4

Graduated Autonomy Levels and Governance Characteristics

Level	Description	Governance Characteristics
Level 0	No autonomy	Rule-based execution; full human control at every step
Level 1	Assisted autonomy	Agent provides recommendations; human approves each action before execution
Level 2	Conditional autonomy	Agent executes routine actions within pre-defined parameters; exceptions and anomalies escalated to human supervisors
Level 3	Bounded autonomy	Agent acts independently within explicitly defined ethical, technical, and institutional limits; continuous HOTL monitoring with exception-based HITL intervention

The above layers provide a detailed explanation on how to implement ethics into autonomous systems. Each layer provides the framework to be able to define what an automated system can do, how much trust should be placed in those decisions, what limitations exist based on context, and where and when human intervention is required.

For instance, if we have an Enterprise AI Agent that is allowed to make approvals of all transaction less than a specific dollar amount then we are establishing both permitted actions, confidence levels, and contextual restrictions. All transactions greater than the specified amount would escalate to human approval.

Layer 2: Ethical Guardrails Layer

The embedding of ethical constraints within the architecture as enforceable system rules supports the implementation of compliance-by-design and responsible AI operationalization (Floridi et al., 2018; OECD, 2019). Examples of ethical guardrails include fairness constraints that monitor disparities in outcomes across different demographic groups, transparency requirements (i.e. explainable decision paths) and consent/privacy bounds.

Ethical guardrails also establish contextual prohibitions in sensitive areas. When an agent experiences a scenario that violates the ethical constraints established by the organization, it must either defer action, initiate human review or revert to a safe default position. Additionally, continuous monitoring of ethical indicators will help prevent "ethical drift" as the agent develops its knowledge base over time.

Layer 3: Human Oversight Layer

Human-in-the-loop (HITL) mechanisms involve real-time human participation in high risk activities, rights impacting decisions, and novel/unusual situations. Human-on-the-loop (HOTL) mechanisms focus on monitoring and auditing the automated activity and enable large-scale human supervision via dashboards, alerts, and regular audits/reviews. As shown in table 4, HITL/HOTL mechanisms require increasing levels of human oversight as the level of automation increases.

Table 5

Oversight Calibration Matrix: HITL and HOTL Intensity by Autonomy Level

Autonomy Level	HITL Intensity	HOTL Intensity
Level 1 — Assisted	High — human approves every action	Moderate — periodic trend review
Level 2 — Conditional	Moderate — human handles exceptions	High — continuous anomaly monitoring
Level 3 — Bounded	Low — exception-based only	Very High — real-time dashboards and audit

Layer 4: Compliance and Accountability Layer

The Compliance and responsibility layer provides the assurance that the behaviour of the AI agent can be traced, audited and proven compliant with laws and regulations. The key building blocks of this layer are:

- Decision logs – these are immutable logs which contain details about each decision made by the agent;
- Explainability artifacts — these provide insight into why the agent arrived at its conclusions;
- Audit trails — this trail links the actions taken by the agent to specific oversight points so that accountability can be tracked;

- Reporting mechanisms — provide an interface for regulators and auditors to report findings. Building these mechanisms into the AI system will ultimately reduce the costs associated with adding Compliance controls after deployment.

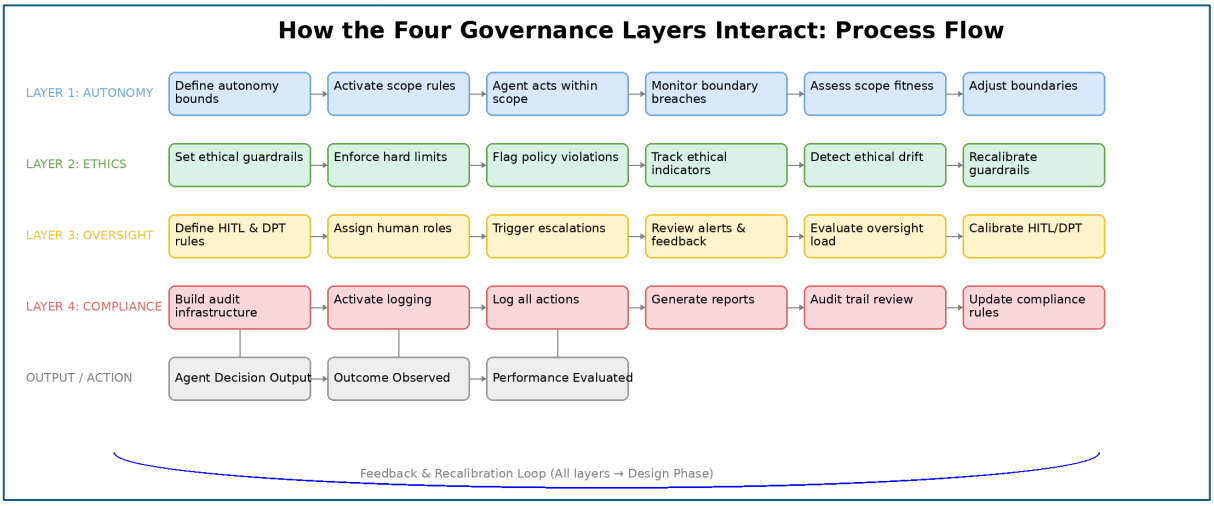
How the Four Layers Interact: Process Flow

Figure 5 depicts the processes involved in implementing all four governance layers. What is illustrated here is that the four layers function together to enable all aspects of the agent transaction cycle to occur concurrently. The autonomy layer defines what the agent has the authority to do. Once defined, the ethics layer verifies if such activity is permissible. During execution, the oversight layer identifies any issues requiring additional escalation.

Finally, the Compliance layer records all results. Throughout the entire life-cycle of the AI system, both positive and negative feedback from both monitoring and review phases will continually feed-back into all four layers. In addition to providing updates to existing boundaries, guard-rails, oversight intensities and Compliance requirements, the feedback also addresses changes in agent behaviour as well as any changes in context.

Figure 5

How the Four Governance Layers Interact: Runtime Execution and Recalibration Process Flow



Source: Author's Inferences

Governance Trade-offs and Failure Modes of Bounded Autonomy

There are several Trade-offs associated with establishing boundaries of autonomy. Too much restriction of autonomy results in less effective operation of a system, higher workloads for humans and lower trust in a system. Conversely, too little constraint exposes an organization to a range of risks including ethical, legal and reputation based risks.

Multi-Agent Systems and Collective Autonomy

Most deployed systems consist of multiple interacting Agents. In some instances collective behavior emerges among agents that cannot be directly attributed to one particular agent. If a governance framework focuses solely on bounded autonomy at an

individual agent Level, there is a high probability that significant systemic effects will be overlooked. Therefore, bounded autonomy needs to be extended beyond individual Agents to collective behaviours involving multiple Agents. This includes developing shared ethical guidelines for agent behaviour, protocols for coordination among Agents and mechanisms for escalating problems that arise among Agents.

Measuring Governance Effectiveness

The success of a governance framework can be measured using several metrics: Frequency and type of escalation; Autonomous vs. Human mediated decisions; Ethics/Compliance violations; Change over time in decision patterns; Latency of human intervention and results of such intervention. These metrics allow organizations to determine if current levels of autonomy are acceptable and if oversight mechanisms are functioning properly. They provide insights to improve governance practices continuously.

Practical Applications and Use Cases

The bounded autonomy governance framework was developed to be flexible enough to apply in a variety of deployment settings. Below are examples of applications in three domains: enterprise systems, public sector infrastructure, critical networked infrastructures and finally, an applied case study demonstrating how all four layers operate together in a single real-world example.

Enterprise AI Agents

AI agents in enterprise settings represents a growing trend towards realizing operational efficiencies through use of AI while simultaneously increasing risks related to governance and reputation (Davenport & Ronanki, 2018; Capgemini Research Institute, 2025). Through bounded autonomy enterprises can achieve operational efficiencies while retaining managerial control. An example of such an AI agent would be an enterprise customer service chatbot that uses conditional autonomy (Level 2), thus allowing it to respond automatically to most routine customer questions while ensuring that it responds equally and fairly regardless of who is making the inquiry.

It would escalate complex or controversial issues to human managers via predetermined escalation rules. Also, as part of its HOTL mechanism, it would monitor itself to detect changes in its response patterns that could indicate performance drift.

Public Sector and Digital Public Infrastructure

Public sector contexts require even greater accountability requirements than private sectors (OECD, 2019; European commission, 2021). As a result, bounded autonomy typically implies lower autonomy levels for AI Agents as well as stricter HITL requirements. The primary role of an AI agent in a public sector setting is usually to assist decision makers by processing very large amounts of information. However, decisions with rights implications always fall under the authority of human officials. Ethics guardrails used in public sector settings tend to emphasize fairness, equity, non-

discrimination and compliance mechanisms maintain auditability/explainability enabling citizens to understand challenges and appeals.

Critical Networked Infrastructures

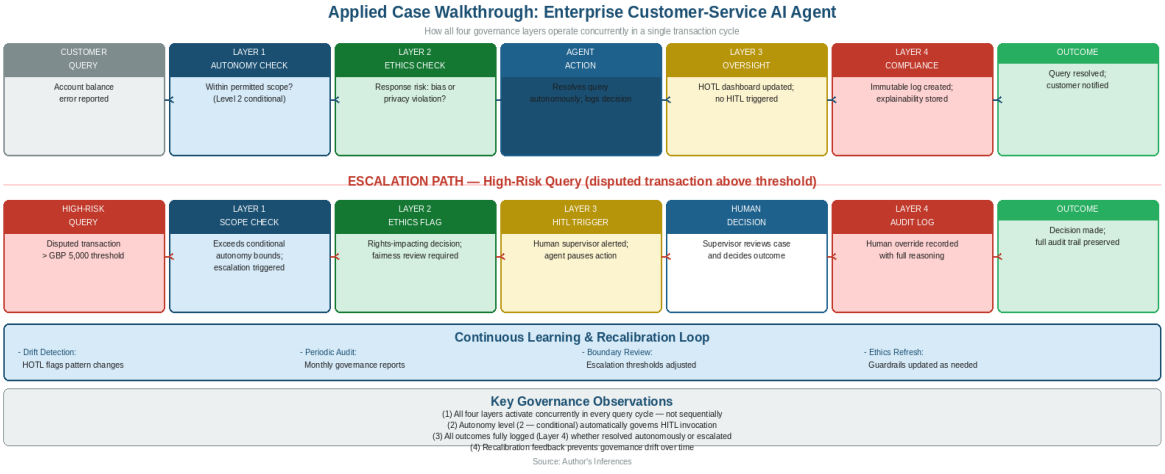
Critical networked infrastructures represent unique challenges due to their inherent complexity as well as their potential impact on society as a whole (Chen et al., 2000). Bounded autonomy plays a major role in balancing risk management with system operability while reducing dependence on human intervention.

Applied Case Walkthrough: Enterprise Customer-Service Agent

To illustrate how this framework works in real-world application, let's take a hypothetical example of a financial services firm implementing a customer service AI agent at Level 2 (Conditional Autonomy) which can handle inquiries on accounts, confirmations of transactions and provide rudimentary dispute resolution. Below is a walkthrough illustrating two different paths through the four levels of governance simultaneously. Figure 6 represents both paths illustrated below.

Figure 6

Applied Case Walkthrough: Enterprise Customer-Service AI Agent — Routine and Escalation Paths



In the first pathway, a customer contacts the firm claiming their account has an error in the reported balance. Layer 1 verifies that the customer's request is within the limits of what the agent is allowed to do based upon its Level 2 Conditional Autonomy (Below Escalation Threshold). Layer 2 determines if there is any bias or potential violation of privacy in the response being provided by the agent. Based upon this determination, the agent provides an autonomous response to the customer's request. Layer 3 updates the HOTL dashboard but does not send an alert to HITL. Layer 4 creates an immutable record of the event and a corresponding explanation artifact. The time to resolve customer issue is less than two minutes and no human interaction occurred.

Implementation Challenges and Organizational Readiness

Implementing bounded autonomy entails an organization having the necessary organizational preparedness in addition to technical capabilities. Organizations often have inadequate governance maturity to implement bounded autonomy. Several common issues include: unclear role definitions, insufficient oversight resources, and lack of Coordination between technical teams and compliance departments.

Implications

The implications of this study align with the broader need for developing governance frameworks that treat AI systems as institutional actors, instead of simply using them as tools in order to address accountability, oversight, and social legitimacy (Floridi et al., 2018; Rahwan, 2018).

Theoretical Implications

This study makes contributions to the field of AI governance by reframing autonomy as a continuum that can be governed, versus a binary attribute. The study integrates sociotechnical systems theory with AI agent research to develop a conceptual model of autonomy that is based upon both institutional context and temporal dimensions. The concept of bounded autonomy provides a basis for further theory development related to systems with agency. The framework expands human-AI interaction research by specifically relating oversight mechanisms to the degree of autonomy of an agent, thereby addressing one major shortcoming of existing HITL and HOTL models.

Managerial and Design Implications

For managers, the framework illustrates the necessity of managing AI governance as a strategic capability. The decisions regarding the delegation of authority over autonomy, designing oversight, and ethical limitations establish boundaries around the potential performance of a system, as well as the risk exposure and reputation of an organization. An organization can minimize its governance costs by incorporating governance considerations into its initial design, and consequently, prevent reactive compliance measures and promote trust proactively. The application case walkthrough illustrated that by configuring the oversight to match the level of autonomy of an agent, most routine interactions do not require any involvement from HITL.

Policy and Regulatory Implications

For policy makers, bounded autonomy represents a practical method to realize broad governance principles. Through aligning the intensity of oversight with the autonomy and impact of an AI system, bounded autonomy facilitates proportionate regulation that promotes innovation while protecting the public interest. The lifecycle nature of the framework is consistent with regulatory philosophies that focus on ongoing monitoring and adaptation, including the European Union AI act's risk-based governance philosophy (European commission, 2021).

Limitations and Future Research

Conceptual and Methodological Limitations

As stated above, this study employs a conceptual DSR methodology, rather than empirically examining the relationship between autonomy, governance and outcomes. While the framework is intended to be technology-agnostic -- enhancing generalizability -- it is possible that specific implementation complexities in architectural designs are overlooked. Additionally, the organizational/institutional perspective emphasized throughout this paper, leaves little room for consideration of user trust/fairness perceptions and/or societal legitimacy related to end-users.

Directions for Future Research

1. Empirical Validation — field studies of AI agent deployments across enterprise and public sector organizations; longitudinal studies evaluating the governance impacts resulting from evolving autonomous systems.
2. Human–AI Interaction Research — perceptions of users concerning agent autonomy/oversight; cognitive load/decision fatigue in HITL systems; determinants of appropriate reliance.
3. Multi-Agent and Ecosystem Governance — accountability/coordination across interacting agents; governance of agent ecosystems spanning multiple organizational and regulatory jurisdictions.
4. Policy and Impact Studies — regulatory sandbox experiments; comparative analysis of governance effectiveness cross-jurisdictionally; integration w/ emerging regulatory frameworks

Conclusion

AI agents represent a major shift in human machine interaction. Unlike previous AI systems designed to assist human decision making, AI agents will execute tasks on their own with little or no human input in a variety of complex and dynamic environments. This represents a huge opportunity and risk for unbounded autonomous action.

The focus of this paper is to identify the major challenges for AI agents and propose a bounded autonomy governance framework to address these issues. As such, the governance framework Includes four key elements: autonomy control, ethical guardrails, calibrated human oversight, compliance-by-design mechanisms. Through its analysis of AI agents as socio-technical actors, the governance framework addresses three significant interdependent issues related to technical innovation, ethical responsibility and institutional accountability.

Three key contributions are made.

- Firstly, the paper transforms our thinking about autonomy from being viewed as either binary or uncontrollable to being viewed as a continuum; therefore, enabling us to view autonomy as a governable property.
- Secondly, the four-layered framework provides organizations with an operational model for designing and operating responsible AI agent systems. The governance framework has been operationalized via a process flow diagram and applied case walkthrough.

- Lastly, the governance framework illustrates how abstract governance principles can be incorporated into real-world systems without stifling innovation.

Therefore, as AI agents continue to expand their presence throughout various industries and public infrastructure domains, the question now facing practitioners, policymakers and researchers is no longer “can machines act independently?” but rather “are institutions governing them adequate?” bounded autonomy presents a principled path forward by enabling AI agents to enhance human capability while maintaining a firm anchor within ethical, regulatory and societal boundaries.

Disclosure of AI Assistance : Generative AI was employed in a selective manner for developing structure and refining language as part of preparing manuscripts. Author is solely responsible for all conceptual development, analysis and reasoning, interpretation and final editing. Use of generative AI is consistent with American Psychological Association (APA) guidelines regarding the ethical use of artificial intelligence.

References

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., & Horvitz, E. (2019). Guidelines for human–AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300233

Benjamins, R., Barbado, A., & Sierra, D. (2019). Responsible AI by design in practice. *AI & Society*, *34*(4), 593–602. https://doi.org/10.1007/s00146-018-0855-7

Capgemini Research Institute. (2025). *AI agents: The next frontier of enterprise transformation*. https://www.capgemini.com/wp-content/uploads/2025/07/Final-Web-Version-Report-AI-Agents.pdf

Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly*, *35*(1), 128–152. https://doi.org/10.2307/2393553

Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. *Harvard Business Review*, *96*(1), 108–116. https://hbr.org/2018/01/artificial-intelligence-for-the-real-world

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. arXiv. https://arxiv.org/abs/1702.08608

European Commission. (2021). *Ethics guidelines for trustworthy artificial intelligence*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Gasser, U., & Almeida, V. A. F. (2017). A layered model for AI governance. *IEEE Internet Computing*, *21*(6), 58–62. https://doi.org/10.1109/MIC.2017.4180835

Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly*, *37*(2), 337–355. https://doi.org/10.25300/MISQ/2013/37.2.01

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75–105. https://doi.org/10.2307/25148625

Hansen, M. T. (1999). The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits. *Administrative Science Quarterly*, *44*(1), 82–111. https://doi.org/10.2307/2667032

Huber, G. P. (1991). Organizational learning: The contributing processes and the literatures. *Organization Science*, *2*(1), 88–115. https://doi.org/10.1287/orsc.2.1.88

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, *93*(5), 1449–1475. https://doi.org/10.1257/000282803322655392

Kahneman, D., & Tversky, A. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, *165*(3), 633–705. https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3

March, J. G., & Simon, H. A. (1958). *Organizations*. Wiley.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 1–21. https://doi.org/10.1177/2053951716679679

OECD. (2019). *OECD principles on artificial intelligence*. OECD Publishing. https://oecd.ai/en/ai-principles

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Piccialli, F., Casolla, G., Cuomo, S., Giampaolo, F., & Iannone, S. (2025). Artificial intelligence agents: Architectures and applications. *Expert Systems with Applications*, *247*, 123456. https://doi.org/10.1016/j.eswa.2025.123456

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

Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, *69*(1), 99–118. https://doi.org/10.2307/1884852

Simon, H. A. (1957). *Models of man: Social and rational*. Wiley.

Simon, H. A. (1972). Theories of bounded rationality. In C. B. McGuire & R. Radner (Eds.), *Decision and organization* (pp. 161–176). North-Holland.

Wooldridge, M. (2021). *A brief history of artificial intelligence*. Flatiron Books.

Zahra, S. A., & George, G. (2002). Absorptive capacity: A review, reconceptualization, and extension. *Academy of Management Review*, *27*(2), 185–203. https://doi.org/10.5465/amr.2002.6587995