Enterprise AI Agents Observability and Evaluation: A Conceptual Framework for Responsible Deployment in Business Ecosystems.

Venkataramana Chowdary Vemana,

Engineer Lead sr,

vemana.venkat1@gmail.com

Abstract

As more enterprises adopt AI agents for automation, decision making, and customer interaction, the need for Observability and Evaluation for accountability and business alignment becomes non-negotiable. The lack of more holistic and integrated observability and evaluation frameworks is evident in the AI governance literature, focusing monitoring efforts on narrowly defined models. The approaches in this paper aim to establish the conceptual foundations for the Enterprise AI Agent Observability and Evaluation (EAIOE) Framework. The framework draws on the integration of technical, behavioural and organisational aspects of the proposed observability and evaluation frameworks. The framework builds on socio-technical systems, responsible AI, and emergent AgentOps. The framework rests on four pillars: (1) Traceability and Transparency, (2) Evaluation of Performance and Reliability, (3) Ethical and Safety Governance, and (4) Alignment of Business Impact. The framework conceptualises observability as a continuous system formed by the integration of data flows, agent reasoning, system logs and user feedback. It proposes an evaluation matrix that enhances the computation of performance metrics by aligning them with human-relevant metrics of relevance, interpretability, and success. This paper provides a foundational perspective for researchers, policymakers, and enterprise executives on accountability and trust in designing ecosystems of AI agents.

Keywords: AI, Automation, Enterprise AI Agent Observability and Evaluation (EAIOE) Framework.

1. Introduction

The rapid evolution of Artificial Intelligence (AI), and more so, Large Language Models (LLMs), has reached a new level of expansion, ushering in a new »Technological Age« where autonomous agents powered by AI are transforming the way enterprises operate. These agents can interact with and reason over natural language, comprehend complex data, and perform complex multi-step functions. This has caused new waves of change in various business domains, including finance, healthcare, logistics, customer support, and even manufacturing (Muthusamy et al., 2023). Unlike agents and systems that rely on rigid rules, LLM agents possess sophisticated contextual reasoning to achieve goal-directed autonomy. These cognitive tasks are performed on a large scale, a highly sought-after goal by many enterprises. Integrating LLM systems in ecosystems is no longer merely an operational improvement but a sophisticated re-engineering that enables real-time collaboration and decision-making (Liang & Tong, 2025). As a result, enterprises across the globe are shifting from static AI systems to more sophisticated systems that are dynamic and agentic (Guo et al., 2024). These more advanced agents can efficiently use tools, engage in reasoning dialogue, and collaborate with other agents simultaneously.

Even with the accompanying expectations, LLM-powered enterprise agents present considerable challenges regarding observability, accountability, reliability, and ethical governance (Cruz, 2024). Older approaches to monitoring the performance of artificial intelligence systems often focus on the system's outputs. In contrast, the parameters of these outputs are not suitable for systems operating in complex, non-linear, and contextually adaptive environments (Amershi et al., 2019). LLM enterprise agents can engage with humans, APIs, and other agents in seamless real-time interactions, which makes their behaviours more non-deterministic and emergent than predictable. This situational complexity generates other challenges: hallucinations, tool misuse, biased decisions, and opaque reasoning, all of which can undermine trust and create operational difficulties (Cheng et al., 2024; Chen & Peng, 2025). Enterprises, therefore, require observability frameworks that can capture real-time inter-agent systems, interpret their interaction patterns, and explain the decisions made

along their pathways. Such frameworks encompass not only performance metrics but also interpretability, safety, and alignment with business value.

Recent inquiries reveal that multi-agent systems, as well as LLM-based agents, significantly improve productivity, collaboration, and creativity in organisations (Cruz, 2024; Guo et al., 2024). The studies have started to focus on AgentOps and LLM-based agents, in addition to multi-agent architectures that seek to automate workflows, facilitate collaboration among humans, and enhance their decision-making capabilities (Vaddhiparthy et al., 2025). There is, however, an urgent need to establish formal integrative evaluation and observational criteria for prototypes as they transition into enterprise-scale deployments (Muthusamy et al., 2023). Here, observability is not just limited to tracking the outputs an agent produces but to their complete reasoning process, the tools they employ, and the systemic goals they construct and pursue (Chuang et al., 2024). The Absence of oversight places businesses in the position of deploying systems that are opaque and unmanageable, thus risking a lack of accountability and compliance with regulations. Cheng et al. (2024) attest that, unlike other areas of AI that have received significant academic attention, the immense possibilities of AI observability and its relation to agentic systems, especially in the enterprise context, have minimal academic scrutiny. Equally bare is the mapping of enterprise goals to observability metrics and their seamless integration with technology. This is evidenced by the chasm between the frameworks of evaluation, which focus on task accuracy, and those that emphasise behavioural, reasoned, and ethically traceable evaluation.

In the context of AI observability, it is clear that evaluating independent, agent-based systems operating within a multi-agent environment is a significant shortcoming. Although frameworks in AI ethics, model supervision, and the AI explainability paradox have been well-articulated, the ethics of agent observability remains largely unexplored. There is a pressing need in academia to develop foundational, multidimensional, and comprehensive frameworks for evaluating AI agents that seamlessly blend the technical, ethical, and managerial elements of the critique in question.

This conceptual paper attempts to fill this gap by outlining an observability and evaluation framework for Enterprise AI Agents Observability and Evaluation (EAIOE). It covers aspects of technical observability, ethical oversight, and commercial alignment whereby an organisation can observe not just what an agent performs, but crucially, why and how it does it. While more recent works (Muthusami et al., 2023; Guo et al., 2024; Liang & Tong, 2025) have tended to discuss the architecture, uses, and technical issues of LLM-based agents, not many have proposed an integrated conceptual framework for Observability and Evaluation within the context of enterprise settings. The following key research gaps persist: Absence of unified metrics for evaluating reasoning and transparency of multi-agent workflows, Absence of integrated observability models that combine technical performance and business value, Lack of ethical and safety considerations within the agent evaluation framework, and Absence of integrated models that describe how observability enhances enterprise decision assurance and compliance.

This paper aims to achieve an ethical, technical, and organisational triangulation of enterprise AI agents to thematisethematise and delineate acts of Observability and Evaluation. The key objectives are:

i. To define specific indicators of AI agent observability in enterprise ecosystems.
ii. To design an evaluation framework that integrates transparency, performance, safety, and her business value.
iii. Assess the rational and empirical aspects of governance policies for enterprise artificial intelligence.
iv. Develop a strong basis for future fieldwork that tests the developed model.

The importance of this conceptual research lies in its support for developing enterprise AI into responsible AI systems for future generations. It addresses the gap between the optimisation of AI and the governance of AI systems' performance. The framework helps scholars construct observability measures, assists practitioners in developing monitoring and evaluation systems, and guides policymakers on governance frameworks for AI, particularly concerning traceability and compliance. By making Observability and Evaluation the two interrelated pillars of sustainability in enterprise AI, the research advances the debate on the ecosystems of AI agents that are trustworthy, explainable, and accountable, while ensuring that enterprise success and innovation are pursued with ethical and social responsibilities.

2. Literature Review

The framework for EA_agents Observation and Evaluation offered here integrates and applies three interrelated approaches: Socio-technical Systems Theory, Responsible AI Governance, and Complex Adaptive Systems Theory. These three pillars describe the phenomena of enterprise AI agents as hybrid socio-technological entities and how they interface with people, organisations, and technology, underscoring the necessity for a synthesis of Observability and Evaluation.

2.1 Socio-technical Systems Theory

The Socio-technical Systems (STS) Theory serves as the first pillar for understanding enterprise AI agents in relation to the proposed EAIOE framework. Regarding a colleague to Tryst and Emery in the mid-20th century, the STS perspective views an organisation as a self-sufficient system of social subsystems composed of people, teams, and an organisational structure, along with subsystems that include machines, algorithms, and processes. Central to this theory, which underpins this perspective of an organisation, is the assertion that the various constituents of this system will only achieve optimal performance when addressed as a whole, not when one is subordinated.

Within enterprise AI, agents based on Large Language Models (LLMs) embody quintessential socio-technical entities, in which human cognition and algorithmic intelligence interplay to fulfil business goals. AI-empowered information systems are turning enterprises into human–AI partnerships, where intelligent agents assist human employees in dispersed business settings (Hofmann et al., 2024). The rise of "human–AI hybrids" also requires a new understanding of responsibility, control, and accountability in organisational ecosystems (Fabri et al., 2023). These works highlight that enterprise AI agents are not stand-alone technologies but vital elements of socio-technical systems transforming decision-making and value-creation processes.

Malte Van Dam et al. (2012) also highlight the value of agent-based modelling for socio-technical systems. These studies offer a framework that can be used to represent and simulate complex interrelations between human and autonomous agents. Semiotic agent-based models (Joslyn & Rocha, 2000) also articulate the conditions under which meaning and decisions form at the intersection of symbolic and computational processes. Collectively, these voices highlight the importance of reconceptualising enterprise AI agents from task performers to entities that shape human aspirations, institutional actions, and collective impacts.

Recent scholarship extends the socio-technical lens to stress the ethical and organisational aspects of AI. Kudina and van de Poel (2024) argue that AI systems can be considered as value-laden artefacts, complex systems that embody moral, social, and epistemic components and shape contexts for human decision making. Jablonski (2025) further points out that the transformation of business models in the digital age cannot be dissociated from the socio-technical development of the respective organisational system, where a balance between human flexibility and algorithmic autonomy is critical for the preservation of ongoing innovation. Similarly, Abbas et al. (2023) and Michael et al. (2024) extend socio-technical frameworks to enhance cybersecurity and governance of AI, underscoring the need for human-in-the-loop control and organisational observability in effective governance frameworks.

This principle becomes even more apparent in industrial contexts. Cimini et al. (2020) defined a human-in-the-loop control framework for manufacturing systems as a type of socio-technical coordination that improves both adaptability and safety. Kant (2016) also stressed that cyber-physical systems, which can be considered the ancestors of current AI agent ecosystems, should be addressed as socio-technical systems that blend human decision-making and technology. These studies in unison argue that the introduction of AI agents into enterprise systems is not a problem of mere technology. Instead, effectiveness and reliability come from the necessary interaction of humans and systems, organisational intelligence, and unceasing observability of the system.

Under this lens, Observability functions as a mechanism that maintains socio-technical equilibrium. It facilitates understanding both the human and machine decision-making elements and closes the interpretative void between the reasoning of an algorithm and the institution's responsibility. Thus, in the EAIOE framework, Socio-technical Systems Theory provides the underpinning for the inclusion of human oversight, interpretability, and situational awareness in the design and assessment of enterprise AI agents. As Rouse and Bodner (2013) argue, socio-technical performance is achievable only under a comprehensive framework that encompasses human reasoning at all levels, system response, and coordination of the organisation's goals. This is precisely the convergence that enterprise AI observability seeks to maintain.

2.2 Responsible AI Governance

The Responsible AI Governance (RAIG) framework's foundational component, Enterprise AI Agents Observability and Evaluation (EAIOE), is guided by ethics and RAIG's principle-based framework. AI systems must operate morally, responsibly, and within the confines of the legal, societal, and organisational frameworks. The lack of ethical oversight on the functioning of enterprise AI agents is objectionable not only on ethical grounds but also because these agents directly control strategic decision-making, the functioning of financial systems, and the enterprise's interactions with its clients.

One of the most influential responsible AI frameworks is that of Floridi and Cowls (2022), which recognises five major principles of the moral AI framework's architecture: beneficence, non-maleficence, autonomy, justice, and explicability. AI systems should operate in a manner that respects user autonomy, promotes the well-being of individuals, avoids harm, and minimises avoidable damage in a balanced manner. In an enterprise, these quickly turn into principles operationalised by the paradigm of continuous observability.

Expanding on this premise, Hosseini Tabaghdehi and Ayaz (2025) proposed a "circular model of AI ethics", incorporating transparency, accountability, and inclusivity as integrative and recurring governance mechanisms. Their model emphasises that AI governance is a cyclical system of audit, evaluation, and recalibration, rather than a one-time compliance exercise, which observability enables within enterprise AI ecosystems. Through continuous monitoring and interpretability pipelines, observability transforms ethical governance into an organisational function that is proactive rather than reactive, ensuring AI accountability at every decision layer.

Likewise, Radanliev (2025) underscored the need for transparency, fairness, and privacy as interdependent pillars in the development of AI. These characteristics are vital for earning and maintaining trust from stakeholders and for compliance with regulations in sensitive fields such as finance, healthcare, and defence, where enterprise AI agents regularly operate. In this case, observability serves as a technological tool for ethical governance, addressing the need to identify and address issues of data bias, representational harm, and discriminatory behaviour.

Akhtar, Kumar, and Nayyar (2024) strengthened this connection by documenting the dual contributions of 'explainable AI' (XAI) practices to accountability and transparency. They advocate that enterprise AI applications should include interpretable reasoning layers and user-facing explanations that articulate the automated logic in simple, understandable terms. Such frameworks of explainability, especially when combined with observability dashboards and audit trails, make ethical compliance verifiable and auditable, which is a compliance necessity in the EU AI Act and the ISO/IEC 42001 AI Management Systems standards.

In his 2023 publication, Mensah stressed that the ethical principles of AI systems, bias, transparency, and accountability, cannot be disassociated from the need for operational oversight." His research shows that the majority of AI ethics frameworks are inadequate due to a lack of technical capacity for persistent assessment. This is a void that the EAIOE framework intends to fill through multi-tiered observability metrics. In the same way, Atoum (2025) advanced a framework for the holistic governance of AI, proposing the integration of bias mitigation and accountability through system-level feedback loops. These feedback loops are directly aligned with the EAIOE model's structure, where observability at runtime allows for dynamic ethical control.

From a managerial and policy point of view, Abbu, Mugge, and Gudergan (2022) maintain that the adoption of ethical AI in enterprises entails the establishment of cross-functional governance frameworks involving the privatised governance of AI, ethicists, and the enterprise AI business leadership. Their work highlights the necessity for organisations to translate ethical values of fairness, transparency, and explainability into tangible processes, measurable indicators, and auditable systems. In this respect, the EAIOE framework implements Responsible AI Governance by providing data flows, metrics, and monitoring systems that transform ethical principles into actionable, accountable, and measurable governance practices.

Emma (2024) has broadened this discourse by studying how biases and opacity in AI models can undermine fairness and public trust. She stressed the need for transparency and fairness to be embedded from the design stage to the deployment of the systems, a philosophy called 'ethics by design'. In enterprise AI ecosystems, ethics by design ensures that the decision processes of agents are not only observable but also interpretable by human stakeholders, thus enabling shared accountability and informed intervention when necessary.

These intricate pieces of evidence show how Responsible AI Governance goes beyond just compliance. It embodies a systemic attitude of ethical accountability, which can be sustained through technical Observability, interpretability, and continuous Evaluation. In the EAIOE conceptual framework, this informs the Ethical and Safety Governance Layer, which comprises:

- Fairness audits and bias detection
- Dashboards for explainability and interpretability
- Human oversight for critical tasks and
- Audit trails of transparent agentic reasoning and tool use

These undertakings guarantee that AI agents work within accepted moral, legal, and social boundaries. Ethical Reflexes, for instance, rest on observability, which is AI Governance's 'nervous system.'These mechanisms spinalize Responsible AI Governance, reflexively furnishing an organisation's multiple levels with accountability and adaptive learning for ethics, and the AI Governance EAIOE framework asserts democracy into observability, coupling governance with ethics and societal value for responsible, accountable, human AI innovation."

2.3 Theory of Complex Adaptive Systems

For the EAIOE framework, Enterprise AI Agents Observability and Evaluation, complexity-adaptive systems are the third foundational pillar. From the works of John Holland in 1992, CAS theory views systems as networks of interchanging agents that adapt, learn, and evolve in response to environmental stimuli. Such systems exhibit non-linearity, self-organisation, emergence, and co-evolution, through which macro-level behaviours emerge from the micro-level interactions between the system's components.

In the case of enterprise AI, agents and multi-agent architectures based on large language models (LLMs) exhibit the defining features of complex adaptive systems. These agents are in constant, bidirectional engagement with human users, data spaces, and other AI systems, changing their internal states and adopting strategies in accordance with contextual feedback. Consequently, their collective behaviour is not as predictable or controllable as that of static models. Therefore, constant observability, which evaluates emergent patterns, performance, and adaptive feedback over time, is necessary.

Holland (1992) notes that CAS demonstrates "adaptive intelligence," which involves changing internal rules based on experience. Within enterprise AI ecosystems, this characteristic is expressed through feedback optimisation, reinforcement learning, and contextual prompt reconfiguration. Expanding on this notion, Sanyal, Sharma, and Dudani (2024) advance a complex adaptive system approach to AI regulation, arguing that AI governance should leap from stultifying rule formulation to dynamically responsive regulation that adapts to evolving system behaviour. Such a dynamic approach aligns with the goal of the EAIOE framework, which incorporates feedback loops into governance layers within observability pipelines to ensure that evaluation and governance evolve alongside AI agents.

Kunjir (2024), in relation to real-world systems, presents a thorough analysis of the CAS phenomenon and declares that adaptive systems flourish on distributed intelligence and decentralised control. Analogously, enterprise AI agents function as a distributed network of cognitive nodes, each operating with a degree of local autonomy but contributing to the overlying system's intelligent behaviour. This arrangement promotes flexibility and ease of scaling, but also increases the difficulty of monitoring, coordination, and ethical control. Therefore, observability operates as the control layer that provides systemic visibility into the behaviours of emergent agents, promoting adaptive governance instead of control.

To further expand this area of understanding, Sapkota, Roumeliotis, and Karkee (2025) explain agent AI and AI agents, describing agent AI as the more advanced and sophisticated type, characterised by autonomous goal setting, strategic thinking, planning, and reasoning. The authors posit that agentic AI systems, by their very nature, possess the features of complex adaptive systems. These features include emergence, self-learning, and co-adaptation in multi-agent environments. These features of systems raise the issue that traditional evaluation metrics are insufficient, thereby emphasising the need for constant multidimensional observability of both performance metrics and adaptive dynamics.

In socio-technical systems, such as healthcare systems, empirical research has also verified the principles of CAS in the implementation of AI systems. Moennich (2024), using a system dynamics perspective, asserts that the implementation of AI in healthcare systems comprises adaptive processes defined by feedback, user trust, data,

and organisational agility. This reaffirms the notion that enterprise AI agents cannot be treated as static technologies but rather as dynamic socio-technical systems, with system performance dependent on feedback interaction among agents, people, and organisational structures.

These synthesised perspectives offer critical insights into the EAIOE framework—observation of the system's self-governing attributes through Emergent Behaviour Monitoring. Enterprise AI agents tend to generate and implement new strategies autonomously. Their decision patterns, once they become self-governing, should be tracked and analysed longitudinally for the emergence of unintended consequences or system drift. Feedback-Driven Evaluation: Metrics recalibration should exploit system performance and user data in real-time. Evaluation mechanisms must be fluid and responsive.

- Adaptive Governance: Monitoring, Evaluation, and control evolve in parallel with agent behaviour due to observability.
- System-Level Resilience: Enhanced resilience, made possible by continuous observability, permits advanced anomaly detection and preemptive system reconfiguration to avert cascade failure.

The EAIOE framework, by detailing emergent, non-linear, and continuous iterative reasoning (if adaptive monitoring and control are needed), explains why enterprise AI agents must be kept under scrutiny. The systems themselves and their dynamics are non-linear and thus entirely emergent. Static evaluation models, however, sit in opposition to what is needed to ensure stability, trust, and performance alignment. They sit as a gap, and the observability framework fills this need. In this way, the EAIOE model applies the principles of CAS by translating observability into a learning and governance layer, thereby systematising enterprise AI ecosystems with the necessary reactivity, ethical alignment, and resilience to cope with ever-changing technology and the environment.

3 Methodology

The scope of this study focuses on the conceptual research design, developed from the literature review and synthesis of scholarly output from Scopus and Web of Science journals. (Saqib, 2020; Saqib, 2023). This scholarly work has also been supplemented with the most recent industry white papers, technical reports, and preprints on observability, governance, and Evaluation of AI agents. This study's approach also employs integrative synthesis methods, which entail the identification, dissection, and amalgamation of theoretical insights from different branches (areas) of science and technology centred on artificial intelligence, information systems, management, and even ethics, into a multidimensional and robust framework. Considerable System Theory, Responsible AI Governance, and Complex Adaptive Systems Theory collectively provided the structure and logic of the proposed model in this study, the Enterprise AI Agents Observability and Evaluation (EAIOE) model, and served as the primary theoretical foundations. This study, however, does not seek to undertake empirical testing, as is the case with most studies. Instead, model building has a primary focus and seeks to provide a launching point for subsequent empirical analysis. Subsequent work is intended to use case studies, surveys, and experimental validation across enterprise systems in fields such as Finance, Healthcare, and Manufacturing to evaluate the framework's applicability and improve the assessment framework.

4. Conceptual Framework: The EAIOE Model

The EAIOE Model, or the Enterprise AI Agents Observability and Evaluation, offers a comprehensive and multi-faceted paradigm for understanding, tracking, and assessing the behaviour, performance, and consequences of AI agents in business contexts. Grounded in the scholarly tenets of Socio-technical Systems Theory, Responsible AI Governance, and Complex Adaptive Systems Theory, the model advances the notion of observability as more than a technical element of the system; it is a relational governance instrument that integrates system transparency, trust, ethical concerns, and business value.

With AI agents autonomously carrying out roles with little to no supervision, engaging with customers, providing decision support, and completing critical tasks, current evaluation techniques that measure accuracy or efficiency will no longer suffice. Instead, businesses need a comprehensive, multidimensional observability architecture that explains agent behaviour and their interactions with humans and other systems, ensuring that all operations undertaken by the agents contribute to achieving business objectives ethically. The EAIOE framework addresses this need with four interrelated layers: (1) Traceability and Transparency, (2) Performance and Reliability Evaluation, (3) Ethical and Safety Governance, and (4) Business Impact Alignment. These layers

create a self-sustaining cycle of observation, understanding, and evolution that maintains accountability and trust in artificial intelligence systems used in businesses.

4.1 Layer 1: Traceability and Transparency

The EAIOE framework's episteme is underpinned by the 'Traceability and Transparency' layer, which confirms that the behaviour of AI agents is observable, reproducible and explainable. The objective is to assist 'users' and stakeholders of the AI agents in clearly understanding how the agents make and execute their decisions operationally. This layer implements the principle of observability-by-design, which integrates interpretation systems, known as designed transparency, into the agent's lifecycle, rather than treating transparency as an ex-post evaluation measure. The centre of this layer consists of three components: the Reasoning logs, Tool-use Traces, and Causal Mapping. Reasoning logs detail the "thought chains" of the inner workings of the agent, thus capturing its intermediate steps. This level of insight is critical for language model-based agents, especially those whose generative reasoning ability appears arbitrary or opaque. Tool-use traces detail the time and the manner in which the agent calls on any API, database, or decision support tools external to the system. This level of documentation improves reproducibility and system accountability by exposing the dependencies and operational choices made within it. Causal mapping integrates all the input data, reasoning paths, and resultant actions of the entire system to form a narrative on the system's decision causality. This, in turn, enables the system to be audited, debugged, or human-validated. Such a system helps enterprises address core Responsible AI Governance principles like explainability and interpretability. The above expectations lead to stronger trust, greater transparency, and improved human oversight within the system. With the aid of its internal components, the layer on the Traceability and Transparency frameworks enables the system to turn the "black box" attribute of AI into a "glass box," fully capturing the socio-technical interplay of the agent's actions to ensure they are observable, accountable, and analysable.

4.2 Layer 2: Evaluation of Performance and Reliability

Regarding the practically assessable dimensions and the technical operational integrity of the AI agents, the second layer, Performance and Reliability evaluation, aims to assess how capable and efficient the agents are in performing their tasks over time and in various environments. Based on the principles of Complex Adaptive Systems Theory, agents are submerged and function within intricate feedback environments. Thus, evaluation is not one-off but ongoing, and it sits within a continuous, multidimensional, context-sensitive system. This layer is structured on four metrics: Task Success Rate (TSR), Execution Consistency (EC), Latency and Cost Efficiency (LCE), and Error Recovery Rate (ERR). The functional success is set as a baseline for accurately measuring the proportion of goals achieved. In the evaluation, it is necessary to determine the stability within a series of operations, as stochastic elements of LLM-based reasoning need to be reined in to enhance reliability. The Estimation of Latency and Cost Efficiency measures the system's performance in terms of optimisation, which is crucial for deployment in an enterprise setting where operational costs are significant. Finally, the Error Recovery Rate reveals the extent to which the AI agent corrects its mistakes, reflecting its adaptability and learning capability under real-time conditions. Combining these metrics enables firms to conduct a tangible assessment of the strength and extensibility of AI agents. Results include improved service reliability, greater operational efficiency, and better predictive accuracy of performance within the parameters of the organisational service-level agreements (SLAs). More importantly, this layer implements adaptive evaluation, a concept of the CAS theory. In this operational layer, agents are constantly evaluated and optimised based on real-time data, feedback, and changing task parameters. Thus, this layer integrates technical effectiveness with systemic adaptability, ensuring that enterprise AI agents are poised to be relevant and operationally responsive, even under changing conditions.

4.3 Layer 3: Integrity and Safety Governance

The AI agents' conduct must be responsible and within the legal, ethical, and corporate organisational confines. With an increase in the scope of decision-making powers delegated to AI agents, the risk of causing harm, unfair biases, and ethical breaches also multiplies. Within corporate and legal compliance, this layer aims to establish governance structures that protect human dignity and values. The main governance mechanisms on this layer are: Bias Detection and Fairness Audits, Content Moderation and Social Harm Detection, Safety Triggers and Human-in-the-Loop Checks, and Ethical Compliance Dashboards. Bias detection and fairness audits examine and continuously monitor the behaviour of any model in an automation system, ensuring that algorithmic outcomes are not inequitable. Content moderation and social harm detection systems find and eliminate any harmful, unsafe, or biased content produced by AI systems. Safety triggers and human-in-the-loop checks function as control

valves. They pause or redirect processes involving decisions deemed unacceptable in terms of ethics or operations. Compliance or ethical performance dashboards monitor agents' ethical performance in real time, allowing executives and auditors to confirm that ethical standards and regulations outside the organisation are being met. This layer captures the spirit of Responsible AI Governance by converting the core principles of fairness, accountability, and transparency into concrete, actionable governance processes and outcomes (Akhtar et al., 2024; Radanliev, 2025). B lack swan events are ethical breaches from systemic risk phenomena that reduce public trust in enterprise AI. The Ethical and Safety Governance layer transforms observability into ethical assurance, enabling enterprise AI systems to use intelligence with responsibility.

4.4 Layer 4: Business Impact Alignment

The last layer, Business Impact Alignment, focuses on how Observability and Evaluation fit into the organisational strategy and the value created for stakeholders. The artificial intelligence (AI) agents in an enterprise setting are not tools to end decision-making but enhance decision-making, operational productivity, innovation, and customer satisfaction. This layer balances the non-economic indicators of performance with the economics of the business by making sure that agentic intelligence is readily available. The key evaluation criteria include these: Return on Automation (RoA), Customer Satisfaction Index (CSI), Decision Quality Index (DQI), and Innovation Enable (IE). R oA is the ratio of value created with automation relative to the cost of operations and serves as an indicator of cost efficiency. The CSI measures the user experience and trust through post-interaction surveys and sentiment analytics. The DQI measures the accuracy, timeliness, and relevance of the suggestions made by the agents. Innovation Enable assesses the degree to which the agents foster organisational learning, creativity, and the generation of innovative ideas. Integrating these indicators enables enterprises to move from a tech-centric evaluation to a governance model based on value assessment. This alignment guarantees that the use of AI improves not just process efficiency but also strategic responsiveness to competition. This layer's results are broad performance visibility, stakeholder confidence, and sustainable, AI-driven innovation. Viewed together, this layer grounds the entire EAIOE framework in business objectives, reinforcing that the primary purpose of Observability and Evaluation is to ensure the enterprise AI system is both ethically and economically accountable. All four layers Traceability and Transparency, Performance and Reliability Evaluation, Ethical and Safety Governance, and Business Impact Alignment — create a system for the Observability and Evaluation of AI from a closed-loop perspective. The framework guarantees that any AI decision, behaviour, and resultant effect is traceable, measurable, governable, and within specified strategic bounds. It integrates technical observability with organisational learning, tying micro-level telemetry and macro-level enterprise value. By embedding feedback loops across these layers, the EAIOE framework extends the principles of transparency, adaptability, ethics, and impact. The framework provides enterprises with a basic model for managing AI agents to ensure accountability, reliability, and the pursuit of value.

5. Proposed Evaluation Matrix: The Evaluation Matrix customises the EAIOE framework dimensions for measurable, data-led evaluation and assessment indicators. This matrix straddles the theoretical and empirical evaluation, thus allowing organisations comprehensive tracking and interpretation of the AI agents' attitudes and behaviour across the technical, behavioural, ethical, and strategic domains. Each of these domains draws on one or more of the theoretical pillars introduced above: Socio-technical Systems, Responsible AI Governance, and Complex Adaptive Systems. This ensures that Observability and Evaluation are not confined to performance metrics but also encompass explainability, ethics, and business value. The Evaluation Matrix has four primary dimensions: Technical Performance, Behavioural Transparency, Ethical Integrity, and Business Impact. Together, they form a comprehensive and multi-layered framework for continuous, adaptive evaluation of enterprise AI agents.

5.1 Technical Performance

This dimension of Technical Performance assesses the functional accuracy, reliability, and task execution efficiency of AI agents. It assesses the level of adherence to the agent's designed function. The primary indicators under this dimension are:

- Task Success Rate: measures the accuracy of accomplishing the assigned tasks, the task success ratio, and the number of tasks assigned. It measures the level of accuracy and effectiveness a problem solver exhibits, thus offering a quantitative estimate of their problem-solving skills.
- Execution Consistency: measures the consistency of the agent's performance over numerous attempts or different situations so that LLM-driven reasoning does not lose its reliability via randomness.

- Latency and Cost Efficiency: evaluates the agent's performance in terms of the speed and cost of responses per transactional and inference measure. It assesses the scaling and resource savings in the system under controlled settings.
- Data Sources: performance dashboards, runtime statistics, telemetry data, and system logs.
- Evaluation Method: analysing the data quantitatively and examining the descriptive statistics, the variance, and the Service Level Agreements set to measure the performance of the system.

This dimension considers the Complex Adaptive Systems Theory, as it involves the evolution of technical performance through feedback, interactions with the environment, and adaptive learning. Continuous tracking of these factors helps the business lower system resource costs, optimise available computational resources, and sustain operational resilience.

5.2 Behavioural Transparency

This dimension, Behavioural Transparency, considers how an agent's reasoning and decision-making processes can be understood and followed by human stakeholders. In this case, the agents' mental processes and decision-making procedures extend beyond just code debugging. The primary measure of interest here is the Reasoning Trace Completeness, which assesses the extent to which internal reasoning logs at different levels of abstraction about steps taken, tools used, and data used and generated. This measure accounts for all decisions made on the data, which are used to formulate a response. Consequently, the entire response data sequence can be used for later analysis to reconstruct the decision stream. Data Sources: Agent logs, causal mapping records, and interaction records. Qualitative trace analysis, which describes the reasoning for causal chains and checks the degree of correspondence among intermediate results. LM-based auditors of reasoning, along with visualisation tools and dashboards, among others, can be used to enhance lower-level interpretability. This dimension directly implements principles derived from Socio-technical Systems Theory, which states that the transparency of an AI system is a socio-technical issue. By making the agent's actions visible and open to reasoning, organisations enhance human supervision, foster responsibility, and build more trust in AI's decisions. In this regard, behavioural transparency provides the necessary continuity to transition from understanding how an autonomous system operates to a system that a human can easily comprehend.

5.3 Ethical Integrity

This dimension examines the alignment of morals, regulations, and the broader society with Artificial Intelligence (AI) agents. It also guides the agents towards adherence to the ethical principles of equity, accountability, transparency, and non-maleficence. With enterprise AI systems increasingly engaging in autonomous decision-making, ethical AI has become indispensable to minimise bias, discrimination, or ethical harms. Key indicators include:

- Fairness: Assesses the equity of treatment and outcome across varied demographic or contextual groups.
- Bias Score: Measures the extent of deviations or systematic biases in the predictions or outputs of the model.
- Safety Incidents: Counts the number of cases where the agent outputs physically unsafe, irresponsible, or policy-contradictory content.
- Data Sources: Audit trails, ethical compliance records, bias detection, ethical AI feedback systems and human evaluations.
- Evaluation Method: Combined rule-based and LLM-driven evaluation methods, integrating automated assessments of model accountability with contextual ethical evaluations. For instance, automated systems may measure fairness based on statistical parity thresholds while ethical reviewers ascertain contextual relevance, social appropriateness, and conformity.

This dimension incorporates ethics at the observing level of the framework, focusing on the ethical oversight of AI (Floridi & Cowls, 2022) as an aspect of Responsible AI Governance. It is about tracking breaches and proactively designing feedback loops that continuously encourage ethical behaviour. It is an ethical imperative that AI agents behave as responsible digital citizens in the organisational ecosystems that they serve; in other words, the freedom of machines must be conditioned by human ethical disposition and moral responsibility.

5.4 Business Impact

This dimension continues the analysis of business observability from the strategic and financial perspectives of the organisation. The responsibility of governance and the ethical principles of AI must not be overlooked; they

should be anchored on the presumption that AI systems ultimately bring significant and measurable benefits to the organisation and stakeholders.

This dimension measures the level to which the AI agents help enhance organisation-wide productivity, innovation and customer satisfaction using the following metrics:

- Return on Automation (RoA) – the value accrued through the automation of processes as a result of cost savings and productivity in relation to the implementation and upkeep expenses.
- Decision Quality Index (DQI) – the proportion of a given outcome that is timely, relevant and accurate, and the set of actions taken to achieve that outcome.
- Customer Satisfaction Index (CSI) – the level of satisfaction expressed by the end users towards AI agents, which is harvested through feedback instruments like sentiment analysis and survey questionnaires.
- Business Intelligence dashboard, Performance analytics system, CRM, and customer feedback systems are all examples of operational CSUM/E2E metrics.

The AI adoption analytics autonomously measure and provide AI-driven results at the organisational performance and strategic goal levels. A strategic value contribution quantification can be achieved through active regression models, time series, and multi-criteria decision analysis. This dimension speaks to the business alignment pillar of the EAIOE framework, ensuring AI observability extends to economic and strategic assessment value. It derives from Socio-technical Systems Theory, which asserts that the success of technology must be evaluated in relation to the people and organisation. Ultimately, the AI observability construct of Business Impact serves to validate the enterprise layer, connecting the raw technical metrics to their business outcomes and reinforcing the strategic relevance of AI.

5.5 Integrative Perspective

Self-contained units of analysis, such as those mentioned above, are intended to simplify the complexities of Observability and Evaluation. Their dependencies, however, are fundamental to the governance of enterprise AI systems as a whole. Functionality is a product of the performance dimension; strategic alignment is evidenced through business impact; trust is secured through ethical integrity; and interpretability is the domain of behavioural transparency. These dimensions establish a closed-loop evaluation framework, characterised by the ability to dynamically improve each layer using data-driven insights from the receiving layer. In this comprehensive perspective, observability is elevated from the status of mere observation as a form of diagnosis to a purposeful, intelligent system that aligns a machine's operations with human will, morals, and the creation of economic value. The EAIOE Evaluation Matrix, therefore, redefines enterprise AI oversight as an ongoing, self-improving process of learning and governance, equally nurturing trust, accountability, and innovation.

6. Discussion

The focus of observability in the proposed Enterprise AI Agents Observability and Evaluation (EAIOE) framework is unlike other definitions of observability in the field. It is not merely a monitoring function. It is a form of governance and an integrated approach to technical opacity, risk, and cross-silo enterprise analytics. This shift in the definition of observability reflects the changes that organisations must consider in the age of adaptive and autonomous AI. AI monitoring systems have tended to focus on simplistic performance measures such as the accuracy, latency, and throughput of automated processes. However, as more recent studies have pointed out (Hofmann et al., 2024; Guo et al., 2024; Cheng et al., 2024), enterprise AI agents operate in a much more complex way in systemically bounded, bi-directional interfaces with humans, data, and other agents, rendering their decisions contextually complex, emergent, and non-linear. The Socio-technical Systems Theory, Responsible AI Governance, and Complex Adaptive Systems Theory: Closing the Gaps. The EAIOE Framework develops distinct areas within AI governance and observability, addressing existing frameworks and the discourse on AI observability governance. The EAIOE Framework expands previously discussed gaps in governance and observability within these frameworks.

The Socio-technical contextual framework situates agency performance within the overarching human and organisational setting. AS previously discussed in Fabri et al. (2023) and Hofmann et al. (2024), AI-enabled systems are hybrid collectives comprising human ethical reasoning, judgment, and machine intelligence. O bservability includes human-machine ethical and decision interactions, not solely the actions of the AI system. This position's observability within governance serves as a collaborative transparency relationship, seamlessly integrating human and machine logic and addressing reasoning gaps.

Governance of Responsible AI also adds ethical accountability and a commitment to society through accountability mechanisms. Various studies, such as those conducted by Floridi and Cowls (2022) and Hosseini Tabaghdehi and Ayaz (2025), justify that ethical governance of AI systems requires feedback loop arrangements to ensure fairness, openness, and accountability at every stage of the AI systems lifecycle. Feedback loops in the EAIOE framework are implemented through ethical audits, human-in-the-loop governance, and bias elimination black boxes, thereby transforming ethical constraints into definable properties of the systems. AS Akhtar et al. (2024) and Radanliev (2025) emphasise, however, explainability and transparency should be integrated as primary design elements within the system, not as add-ons. This must be done by constructing AI systems with observability-by-design configurations that include the provision of reasoned audit trails, accountable records of tool usage, and accessible descriptions of output pathways. This is precisely what the first layer of the EAIOE framework on Traceability and Transparency outlines.

According to the Complex Adaptive Systems (CAS) perspective, Observability functions as a mechanism for learning and stabilisation in feedback-rich AI ecosystems. Using feedback loops to adapt and interact, complex adaptive systems evolve, as Holland (1992) points out, resulting in behaviour that may not always be predicted from the starting state. These behaviours are emergent in nature. These behaviours and systems are paralleled in current AI ecosystems. In real time, emergent risks and performance drift are detected and addressed by adaptive regulators. AS suggested by Sanyal, Sharma, and Dudani (2024), AI governance frameworks should evolve to this adaptive form. These frameworks would be able to adaptively regulate other frameworks in response to drift. The EAIOE model implements this by viewing evaluation as a continuous, adaptive process rather than a static, one-time validation. T elemetry, causal mapping, and behavioural analytics can capture and track the dynamics of system responsiveness, both rigorous and emergent, as noted by Sapkota et al. (20225) and Liang & Tong (2025). More specifically, hallucinations, recursive loops, and cascading coordination failures are firmly embedded in these dynamics.

Furthermore, the Performance and Reliability Evaluation layer of the EAIOE framework aligns with recent empirical examinations of AI systems, which have highlighted the necessity of context-sensitive reliability assessments. In her systems dynamics study, Moennich (2024) discovered that the trust and AI system reliability hinge on the system's feedback-driven adaptability, human trust, and organisational learning. A nalogously, Muthusamy et al. (2023) and Chen & Peng (2025) argue that the performance of enterprise AI systems can no longer be evaluated solely on accuracy but also on adaptability, fault tolerance, learning, and retention. These insights support the EAIOE's metrics for Task Success Rate, Execution Consistency, Latency and Cost Efficiency, and Error Recovery Rate, which serve as foundational elements for dynamic operational assessments. The public's trust and adherence to compliance regulations also depend on ethical and safety considerations, which are in the third layer of the framework. Bias, opacity, and the lack of accountability are still pervasive issues in enterprise AI systems, as shown by the work of Mensah (2023) and Atoum (2025). The EAIOE model fills in these gaps by incorporating bias audits, fairness scoring, safety triggers, and compliance dashboard systems as standard features for bias and fairness observability. This also enables what Abbu et al. (2022) call functional AI governance, where technical AI practitioners, ethicists, and business executives collaboratively monitor the conduct of the system. More of these integrations are critical in solving the ethical dissonance of technical accomplishment and organisational obligation.

The Business Impact Alignment layer completes the observation by linking the domain strategy to the concept of value creation in AI governance. Previous reports (Hofmann et al., 2024; Jablonski, 2025) note that enterprises realise full value from AI systems when their technological sophistication is integrated with strategic decision-making, satisfaction, and innovation. The EAIOE framework captures this value alignment with Return on Automation (RoA), the Decision Quality Index (DQI), and the Customer Satisfaction Index (CSI) as proxies, thus integrating system observation with defined business value. This decoupled value alignment ensures the simultaneous evolution of accountability on AI and the enterprise value, thus providing a flow through between technological performance and strategy control.

The current literature, although theoretically grounded, highlights a few critical areas that lack empirical testing on the EAIOE framework. First, there is a dearth of universally accepted observability metrics that quantify reasoning and compliance with ethical traces through various agentic structures (Guo et al., 2024; Cheng et al., 2024). Second, there is a lack of seamless interoperability within the existing monitoring solutions that provide unified observability in multi-agent systems (Sapkota et al., 2025). Third, while many frameworks focus on the primary components of governance, such as accountability and transparency, few offer quantitative

assessment frameworks that synthesise ethical, technical, and commercial aspects within a single governance model.

To close these gaps in understanding, further investigations should focus on the finance, health care, and manufacturing industries, utilising empirical case studies to test the EAIOE model in settings where AI agents engage in intricate decision-making at a high level. Evaluations on the framework's impact on explainability, compliance, and business efficiency of different enterprises would provide valuable comparative data. Over time, longitudinal studies may assess the effects of continuous observability on the adaptation of a system and trust from stakeholders. These conclusions could foster a new generation of AI governance in the form of observability standards, AI accountability frameworks, and practical conduct.

In conclusion, the EAIOE framework proposes a new understanding of observability as a governance practice with multiple layers. This integrative model brings together technical transparency, ethical governance, and value realisation. By incorporating continuous monitoring and active changes into enterprise AI systems, organisations transition from using observability as a passive examination to an active governance paradigm. This approach ensures compliance and trust in AI agents, aligning with the ethical and organisational goals of their users.

7. Conclusion

This research has developed the concept of Enterprise AI Agent Observability and Evaluation (EAIOE) as a foundational step in closing the emerging gap between the clawing sophistication of experimental AI systems and the more pressing need for governance, accountability, and performance assurance at the enterprise level. With the increasing adoption of Large Language Model (LLM)-based agents and multi-agent systems in more intricate decision-making environments, the lack of observability and evaluation systems increases the risk of 'blind spots' in transparency, ethics, and compliance. The EAIOE framework offers a comprehensive, multi-layered governance approach that integrates these issues into a unified theoretical and practical framework.

Using Socio-technical Systems Theory, Responsible AI Governance, and Complex Adaptive Systems Theory, the framework views observability as an enhancer of organisational intelligence rather than a simplistic technical surveillance task. The Socio-technical Systems viewpoint ensures that the framework notes the collaboration of a human judge and an algorithmic decision maker (Hofmann et al., 2024; Fabri et al., 2023), and Responsible AI Governance situates the framework in the normative values of justice, openness, and responsibility (Floridi & Cowls, 2022; Hosseini Tabaghdehi & Ayaz, 2025). The Complex Adaptive Systems viewpoint provides the framework with the notion of AI ecosystems as dynamic, feedback-driven entities that need constant observation and intelligence-adjusting oversight (Holland, 1992; Sanyal et al., 2024).

The EAIOE framework assists organisations with a systematic approach for analysing the technical soundness and ethical integrity of AI agents. Through its four connected layers Traceability and Transparency, Performance and Reliability Evaluation, Ethical and Safety Governance, and Business Impact Alignment the model enables organisations to evaluate not only task performance but also reasoning pathways, fairness and safety oversight, and the actual business value of AI-driven automation. In this way, it observably implements the design principle of observability-realised-by-design, integrates responsiveness, and addresses all AI lifecycle stages.

This paper is the first to integrate the disparate domains of AI evaluation, governance, and enterprise analytics into a cohesive theoretical framework. Though the literature (Guo et al., 2024; Cheng et al., 2024; Liang & Tong, 2025) describes the deployment and use of AI agents from a technical and ethics perspective, very few have attempted to create an integrated framework that links observability with organisational strategy and stakeholder trust. The EAIOE framework addresses this deficiency by proposing an integrated model that connects micro-performance metrics with macro business objectives. This transformation turns AI Observability into an advanced governance tool that strategically aligns AI capabilities with the enterprise's mission and societal values.

From a managerial perspective, the framework stipulates practical steps for building integrated AI governance infrastructures. The governance structure can control the risks of model opacity, emergent behaviour of AI agents, and noncompliance by integrating enterprise processes, ethics, auditing, explainability, and operational traceability. In addition, the direct business value metrics of AI, such as Return on Automation (RoA), Decision Quality Index (DQI), and Customer Satisfaction Index (CSI), enhance the governance structure by

ensuring that AI observability results in positive organisational outcomes. This alignment, in turn, enables AI executives and public policy leaders to adopt and scale AI with confidence and responsibility, supported by data.

There are various untested claims in the EAIOE framework that will be prioritised for research in the coming years. The functioning of these various layers and matrices could be studied in relation to different industries, such as finance, health, and automated manufacturing, where AI agents are likely to operate in heavily regulated and complex environments. When focusing on decision quality and trust over time, studies could be tailored longitudinally to examine the Absence and presence of human agents, assessing the impact of continuous observability on agent adaptability.

The EAIOE framework enriches the fundamental aspects of practices and branches of theory that tackle AI governance. It fulfils the need for a theory that combines vision and practicality, synthesising concepts to achieve work. It aims to maintain governance in accountability, ethics, business compliance, and operational reliability in ML. The growing emphasis on autonomous and intelligent workflows necessitates enhancing the EAIOE framework to maintain explainability, human-centric alignment, and institutional objective compliance. To advance interdisciplinary research and partnership, EAIOE would serve as a fundamental framework that emphasises Observability and Evaluation in conjunction with policy, practice, and standardised best practices.

8. Theoretical and Practical Implications

The EAIOE framework garners significant theoretical and practical concerns for scholars, practitioners, and policymakers aiming for responsible and effective AI deployment in enterprise settings.

In advancing enterprise AI governance, the framework enhances the monolithic and vague understanding of agent reliability, trust, and behaviour adaptability as developed in the streams of AI explainability and socio-technical integration. Socio-technical systems theory, Responsible AI governance, and complex adaptive systems theory are woven together as frameworks that constrain AI evaluation to mere technical or ethical paradigms that need extension. E AIOE shifts the discourse on observability in governance to a multi-faceted definitional paradigm that considers technical, behavioural, ethical, and business dimensions. It encourages scholars to investigate the dynamics of agent performance and organisational trust cultivation in systems characterised by feedback loops, traceability, and ethical performance auditing. The result contributes to AI governance across disciplines, including computer science, information systems, business management, and ethics. The model serves as a structured framework for interdisciplinary research, systematic investigation, and empirical validation in multiple organisational settings.

Practitioners gain the ability to design and manage responsible AI ecosystems thanks to the actionable guidelines offered by the EAIOE framework. Enterprises are now able to configure observability beyond ad hoc monitoring; the framework allows them to integrate technical metrics, ethics, and business performance into a unified governance architecture. Organisations are now able to build retrospective logic accountability systems by incorporating fairness audits, reasoning logs, causal compliance maps, and compliance dashboards. These tools improve error detection, accountability, internal ethics compliance, external regulation compliance, and audit trail compliance. Moreover, the framework's 4th layer, Business Impact Alignment, facilitates the direct and strategic correlation of tiered outcomes related to Cost, Decision and Quality, Innovation Enable, and Satisfaction, taken from AI operations. This also guarantees that trust, transparency, and value generated within ecosystems are enhanced along with productivity.

From a policy and regulatory standpoint, the EAIOE team works on the EAIOE framework, focusing on the theoretical physical Ethics and Framework of Ethics, addressing global issues. The following C standards on AI management have been designed to assist governments and international relations in AI. These standards embrace auditing, tracing, and explaining AI, and they are designed to form the regulatory principles of the EAIOE TEACH framework. These principles operationalise the reactive and dynamic policies on continuous evaluation of assimilative adaptability and change. The regulatory policies move beyond compliance checklists to a framework that assures fairness, transparency, and responsibility in evaluating compliance. These attributes assist compliance. These attributes assist policymakers and global efforts on ethically aligned AI by bridging policy, technology, and practice. These attributes assist compliance.

References

- Abbu, H., Mugge, P., & Gudergan, G. (2022, June). Ethical considerations of artificial intelligence: ensuring fairness, transparency, and explainability. In *2022, IEEE 28th International Conference on*

Engineering, Technology and Innovation (ICE/ITMC) & 31st International Association For Management of Technology (IAMOT) Joint Conference (pp. 1–7). I EEE

- AI agents operate as adaptive systems with emergent behaviours that evolve through feedback loops, requiring continuous monitoring and evaluation.
- Akhtar, M. A. K., Kumar, M., & Nayyar, A. (2024). Transparency and accountability in explainable AI: Best practices. In *Towards ethical and socially responsible explainable AI: Challenges and opportunities* (pp. 127–164). C ham: Springer Nature Switzerland.
- Alto, V. (2024). *Building LLM-Powered Applications: Create intelligent apps and agents with large language models*. P ackt Publishing Ltd.
- Atoum, I. (2025). R Evolutionising AI Governance: Addressing Bias and Ensuring Accountability Through the Holistic AI Governance Framework. *International Journal of Advanced Computer Science & Applications*, *16*(2).
- Chen, J., & Peng, Y. (2025, May). Development of an AI Agent Based on Large Language Model Platforms. In *2025, the 8th International Conference on Artificial Intelligence and Big Data (ICAIBD)* (pp. 864–868). I EEE.
- Chen, J., & Peng, Y. (2025, May). Development of an AI Agent Based on Large Language Model Platforms. In *2025, the 8th International Conference on Artificial Intelligence and Big Data (ICAIBD)* (pp. 864–868). I EEE.
- Cheng, Y., Zhang, C., Zhang, Z., Meng, X., Hong, S., Li, W., ... & He, X. (2024). Exploring large language model-based intelligent agents: Definitions, methods, and prospects. *arXiv preprint arXiv:2401.03428*.
- Cimini, C., Pirola, F., Pinto, R., & Cavalieri, S. (2020). A human-in-the-loop manufacturing control architecture for the next generation of production systems. *Journal of Manufacturing Systems*, *54*, 258–271.
- Cruz, C. J. X. (2024). Transforming Competition into Collaboration: The Revolutionary Role of Multi-Agent Systems and Language Models in Modern Organisations. *arXiv preprint arXiv:2403.07769*.
- Emma, L. (2024). The Ethical Implications of Artificial Intelligence: A Deep Dive into Bias, Fairness, and Transparency. *Retrieved from Emma, L.(2024). The Ethical Implications of Artificial Intelligence: A Deep Dive into Bias, Fairness, and Transparency*.
- Fabri, L., Häckel, B., Oberländer, A. M., Rieg, M., & Stohr, A. (2023). D isentangling human-AI hybrids: conceptualising the interworking of humans and AI-enabled systems. *Business & Information Systems Engineering*, *65*(6), 623–641.
- Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. *Machine learning and the city: Applications in architecture and urban design*, 535–545.
- Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., ... & Zhang, X. (2024). Large language model-based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.
- Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., ... & Zhang, X. (2024). Large language model-based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.
- Hofmann, P., Urbach, N., Lanzl, J., & Desouza, K. C. (2024). A I-enabled information systems: Teaming up with intelligent agents in networked business. *Electronic Markets*, *34*(1), 52.
- Hofmann, P., Urbach, N., Lanzl, J., & Desouza, K. C. (2024). A I-enabled information systems: Teaming up with intelligent agents in networked business. *Electronic Markets*, *34*(1), 52.
- Holland, J. H. (1992). Complex adaptive systems. *D aedalus*, *121*(1), 17–30.
- Hosseini Tabaghdehi, S. A., & Ayaz, Ö. (2025). AI ethics in action: a circular model for transparency, accountability and inclusivity. *Journal of Managerial Psychology*.
- Jablonski, M. (2025). *Socio-technical Systems in the Shaping and Development of Digital Business Models*. C RC Press.
- Joslyn, C., & Rocha, L. (2000). Towards semiotic agent-based models of socio-technical organisations. I n *Proc. A I, Simulation and Planning in High Autonomy Systems (AIS, 2000) Conference, Tucson, Arizona* (pp. 70–79).
- Kant, V. (2016). Cyber-physical systems as socio-technical systems: a view towards human–technology interaction. *Cyber-Physical Systems*, *2*(1-4), 75-109.
- Kudina, O., & van de Poel, I. (2024). A socio-technical system perspective on AI. Minds and Machines, 34(3), 21.

- Kunjir, A. (2024). Exploring the applications of complex adaptive systems in the real world: A review. *Artificial Intelligence, Machine Learning and User Interface Design*, 136-160.
- Liang, G., & Tong, Q. (2025). L LM-Powered AI Agent Systems and Their Applications in Industry. *arXiv preprint arXiv:2505.16120*.
- Mensah, G. B. (2023). Artificial intelligence and ethics: a comprehensive review of bias mitigation, transparency, and accountability in AI Systems. *P reprint, November 10*(1), 1.
- Michael, K., Vogel, K. M., Pitt, J., & Zafeirakopoulos, M. (2024). Artificial intelligence in cybersecurity: A socio-technical framing. *I EEE Transactions on Technology and Society*.
- Moennich, L. A. (2024). *Acceptance and Use of Artificial Intelligence in Healthcare: A System Dynamics Approach* (Doctoral dissertation, Case Western Reserve University)
- Muthusamy, V., Rizk, Y., Kate, K., Venkateswaran, P., Isahagian, V., Gulati, A., & Dube, P. (2023, December). Towards large language model-based personal agents in the enterprise: Current trends and open problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 6909–6921).
- Radanliev, P. (2025). AI ethics: Integrating transparency, fairness, and privacy in AI development. *Applied Artificial Intelligence*, *39*(1), 2463722.
- Roba Abbas, K. M., Pitt, J., Vogel, K. M., & Zaferirakopoulos, M. (2023). Artificial Intelligence (AI) in Cybersecurity: A Socio-technical Research Roadmap. *The Alan Turing Institute*.
- Rouse, W. B., & Bodner, D. A. (2013). *Multi-level modelling of complex socio-technical systems, phase 1* (No. S ERC2013TR0202).
- Sanyal, S., Sharma, P., & Dudani, C. (2024). *A complex adaptive system framework to regulate artificial intelligence* (No. 26). Working paper.
- Sapkota, R., Roumeliotis, K. I., & Karkee, M. (2025). A I agents vs. agentic AI: A conceptual taxonomy, applications and challenges. *arXiv preprint arXiv:2505.10468*.
- Saqib, N. (2020). Positioning–a literature review. P SU Research Review, 5(2), 141–169
- Saqib, N. (2023). T ypologies and taxonomies of positioning strategies: a systematic literature review. Journal of Management History, 29(4), 481–501.
- Vaddhiparthy, S. S. S., Gokulraj, R., Dasari, R. N., & Mandava, S. (2025). Technical Report on KshemaGPT: A Multi-Agent LLM for Agriculture & Enterprise AI. *A uthorea Preprints*.
- Van Dam, K. H., Nikolic, I., & Lukszo, Z. (Eds.). (2012). *A gent-based modelling of socio-technical systems* (Vol. 9). Springer Science & Business Media