# AI Agents: Governance, Evaluation, and State of the Art

Your Name
*Department of Computer Science*
*Your University*
Your Email

***Abstract*—This report synthesizes insights from three seminal papers on AI agents: (1) Governing AI Agents (Kolt), (2) AI Agents That Matter (Kapoor et al.), and (3) AI and Agents: State of the Art (Alonso). It explores the transition from language models to autonomous agents, the benchmarking and evaluation challenges in current research, and the broader vision of agents as the foundation for future artificial intelligence systems. The analysis highlights governance, economic, technical, and design perspectives to provide a comprehensive understanding of AI agents in theory and practice.**

***Index Terms*—AI Agents, Governance, Benchmarking, Multi-agent Systems, Artificial Intelligence**

## I. Introduction

AI research is undergoing a paradigm shift. Beyond static tools such as language models, artificial intelligence is now embodied in autonomous agents capable of planning, decision-making, and execution with minimal human input. This report integrates three research perspectives to provide an overview of governance frameworks, benchmarking challenges, and state-of-the-art advances in AI agent design.

## II. Literature Review

The literature spans legal, technical, and conceptual discussions of AI agents. Kolt's work situates AI agents within principal-agent theory and agency law, identifying challenges of authority, loyalty, and accountability. Kapoor et al. critique the benchmarking practices that overemphasize accuracy at the expense of cost, reproducibility, and real-world utility. Alonso reflects on the conceptual foundations of agents, emphasizing autonomy, flexibility, learning, and social cooperation as defining features of intelligent systems.

## III. Comparative Analysis

Together, these works underscore complementary dimensions of the AI agent discourse. Kolt provides a governance framework rooted in law and economics, Kapoor et al. highlight empirical pitfalls in current evaluation practices, and Alonso situates agents within a broader philosophical and technological trajectory. The convergence of these perspectives reveals tensions between rapid deployment and careful regulation, as well as between theoretical ambition and practical implementation.

## IV. Challenges

Major challenges include information asymmetry between users and agents, risks of overfitting in benchmarks, lack of standardization in evaluation, limited interpretability, and unresolved issues of liability. There are also engineering obstacles such as agent adaptability, design methodologies, and interoperability across domains.

## V. Future Directions

Future research should integrate governance frameworks with empirical benchmarking standards, ensuring AI agents are not only high-performing but also trustworthy and cost-efficient. There is also a need for unified design languages, robust evaluation protocols, and interdisciplinary collaboration across law, economics, computer science, and ethics.

## VI. Conclusion

AI agents represent a defining moment in the evolution of artificial intelligence. The synthesis of governance-oriented, evaluation-focused, and conceptual perspectives demonstrates both the promise and risks of agentic AI. A balanced approach—grounded in law, empirical rigor, and technical innovation—will be essential to ensure agents develop as beneficial and accountable systems.

## References