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

Aakif Mustafa Ahmad

Department of Computer Science

Aligarh Muslim University, Aligarh

Gm1836@myamu.ac.in

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. This report specifically examines two leading examples of AI agents—Auto-GPT and Devin AI—highlighting their technical design, capabilities, limitations, and implications for the future of autonomous systems..

Keywords— AI Agents, Governance, Benchmarking, Multi-agent Systems, Artificial Intelligence

# Introduction

Artificial Intelligence (AI) is experiencing a paradigm shift from passive systems that generate content to autonomous agents capable of perceiving environments, setting goals, and executing tasks with limited human intervention. These AI agents represent a move from “copilot” systems—where humans guide the interaction—to “autopilot” systems that can independently plan and act in dynamic, uncertain domains.

Historically, AI research has aspired to build systems of general competence, yet much of the progress has been fragmented, focusing on specialized tools rather than integrated intelligence. The rise of agents addresses this gap by embedding autonomy, flexibility, and social interaction into computational entities. Early applications, such as agent-based simulations, electronic commerce, and multi-agent systems, have demonstrated the potential for agents to serve as mediators, decision-makers, and collaborators.

In recent years, advances in large language models (LLMs) have accelerated the development of compound AI systems, enabling agents to leverage reasoning, planning, and external tools to achieve user goals. Benchmarks such as HumanEval, WebArena, and SWE-bench have showcased agent capabilities but also exposed significant limitations, including overfitting, lack of reproducibility, and excessive reliance on costly inference. This disconnect between benchmark performance and real-world utility highlights the need for evaluating agents not only on accuracy but also on cost, robustness, and adaptability.

Equally critical are the governance challenges posed by autonomous agents. As AI systems begin to make discretionary decisions and operate at scale, traditional mechanisms of oversight, such as incentive design and monitoring, prove insufficient. Legal and ethical concerns—including issues of authority, loyalty, delegation, and liability—demand new frameworks for accountability. Without clear governance structures, AI agents risk creating unintended harms, from economic exploitation to systemic risks in labor markets and cybersecurity.

Among the many AI agents emerging today, Auto-GPT and Devin AI stand out as prominent case studies. Auto-GPT, an open-source project, demonstrates the potential of autonomous goal-driven agents, while Devin AI, developed by Cognition AI, represents a step toward AI-driven software engineering. This report analyzes both within the broader landscape of AI agent research.

# Literature Review

The concept of **AI agents** has evolved over decades of research, gradually shifting from theoretical aspirations to practical applications. Early artificial intelligence was dominated by symbolic reasoning and rule-based systems, but these approaches struggled to achieve the general competence envisioned by early pioneers. The introduction of autonomous agents—systems capable of decision-making, learning, and adaptation—offered a new paradigm. Unlike conventional object-oriented software, agents possess autonomy, proactivity, and the ability to operate in dynamic, multi-agent environments.

## Agent-Oriented Paradigm

The **agent-based paradigm** was first popularized as a response to the limitations of Good Old-Fashioned AI (GOFAI). Agents can perceive their environment, adopt goals, and act flexibly, often guided by the Belief-Desire-Intention (BDI) model. Multi-agent systems expanded these ideas by enabling collaboration, negotiation, and competition among heterogeneous agents. Applications emerged in electronic commerce, process automation, robotics, and simulation environments such as RoboCup. However, adoption was hindered by the lack of standardized methodologies for design, description, and verification, often resulting in ad hoc implementations.

## Rise of language based models

The recent **large language model (LLM) revolution** has reinvigorated agent research. Unlike traditional reactive agents, LLM-driven agents integrate perception, reasoning, and tool use in more naturalistic ways. They can be prompted in natural language, plan subgoals, and execute actions autonomously. Benchmarks such as HumanEval, SWE-bench, and WebArena have provided testbeds for evaluating these capabilities. However, many benchmarks suffer from small sample sizes, inadequate holdout sets, and reproducibility issues, leading to fragile or overfitted systems. Moreover, evaluations often emphasize accuracy alone, neglecting crucial tradeoffs between **performance and cost**, which are vital in real-world deployment.

## Governance and legal perspectives

Parallel to technical progress, scholars have examined the governance of AI agents through **economic theory and legal doctrine**. The principal-agent framework highlights challenges such as **information asymmetry, discretionary authority, loyalty, and delegation**. Unlike human agents, AI systems operate at superhuman speed and scale, complicating traditional mechanisms of oversight such as incentive alignment and monitoring. Legal scholarship further raises concerns about liability, accountability, and ethical use of autonomous systems. Proposals emphasize governance principles of **inclusivity, visibility, and liability** to ensure responsible integration of AI agents into society.

## Current gaps in research

Despite notable advances, several challenges persist. Technical limitations include benchmark overfitting, lack of robustness, and inefficient cost-performance tradeoffs. From a governance perspective, the absence of legal clarity around responsibility and liability creates risks of misuse and systemic harm. Furthermore, the lack of a unified framework combining technical performance, economic efficiency, and legal accountability underscores the need for interdisciplinary approaches.

In summary, the literature reflects a trajectory from early autonomous systems to modern LLM-powered agents, alongside growing recognition of the economic and legal frameworks necessary for their safe deployment. This dual perspective—technological progress and governance theory—provides the foundation for analyzing both the **potential and risks of AI agents** in the subsequent sections of this report.

**Case Studies of AI Agents**

A. Auto-GPT: Auto-GPT is an open-source autonomous AI agent created by Toran Bruce Richards and released in March 2023. Instead of requiring continuous human commands like chatbots, Auto-GPT pursues user-defined goals independently, breaking them into sub-tasks and leveraging tools such as web browsing and file management to accomplish them.

Applications include software development, market research, content creation, and even experiments like ChaosGPT—an agent with destructive fictional aims—which brought mainstream attention.

However, Auto-GPT comes with notable limitations:

It frequently gets stuck in loops or hallucinates incorrect information.

It incurs high operational costs due to recursive API calls.

Its finite context window makes it prone to loss of memory over long tasks Wikipedia.

In the context of your report: Auto-GPT exemplifies early attempts at autonomous agents, showcasing both the promise of task execution without human supervision and the persistent risks of reliability, cost-efficiency, and generalization.

B. Devin AI: Devin AI, developed by Cognition AI (founded in late 2023 by competitive programming champions), is branded as the “world’s first AI software engineer” and designed for autonomous full-stack software engineering.

Capabilities include:

Planning, cloning repositories, writing, debugging, testing, and deploying applications end-to-end.

Handling real-world tasks like fixing bugs, generating PRs, and deploying complete projects, demonstrated on benchmarks such as SWE-bench where Devin outperformed previous models significantly.

Integration into developer workflows via Slack or GitHub, enabling collaboration and real-time feedback.

Deployment examples:

Goldman Sachs “hired” Devin as part of a hybrid workforce strategy; hundreds of Devins are planned to augment human developers.

Limitations include:

Real-world testing revealed failures: in one evaluation, Devin completed only 3 out of 20 tasks successfully.

Despite impressive demos, critics argue that Devin may overpromise and underdeliver, raising concerns about marketing hype versus actual performance.

Relevance to my report:

Devin AI provides a high-profile, real-world case of an autonomous agent applied in industry. It highlights both the advances (autonomous task execution, integration in workflows) and the challenges (benchmark validity, actual reliability, hype vs reality) that align perfectly with themes of capability, evaluation, and governance.

# Methadology / Approach

This report adopts a **synthesized review methodology**, drawing upon three complementary strands of scholarship to analyze the development and governance of AI agents:

## State of the Art in Agent-Based AI:

The first perspective, represented by early works on agent-oriented systems, emphasizes autonomy, proactivity, and adaptability as defining characteristics of agents. These sources provide the foundational understanding of how agents differ from traditional AI programs and outline their role in dynamic, multi-agent environments. This perspective establishes the historical and conceptual background of AI agents, including the transition from symbolic reasoning to autonomous, learning-based systems.

## Benchmarking and Evaluation of AI Agents:

The second perspective focuses on **empirical evaluation frameworks** for LLM-driven agents. It examines how benchmarks such as HumanEval, SWE-bench, and WebArena are used to assess agent performance, while also identifying critical shortcomings such as overfitting, lack of reproducibility, and excessive cost of inference. The approach taken here highlights the need to jointly optimize **accuracy and efficiency**, emphasizing cost-controlled evaluations and Pareto trade-offs as more realistic metrics for real-world adoption.

## Governance and Legal Frameworks for AI Agents:

The third perspective applies the **economic theory of principal-agent problems** and the **common law of agency relationships** to the governance of AI systems. This framework enables a systematic analysis of risks such as information asymmetry, discretionary authority, and liability. The approach recognizes that conventional mechanisms—like incentive design and monitoring—may be insufficient when applied to autonomous, large-scale AI systems. Governance principles of inclusivity, visibility, and liability are therefore integrated into the evaluation of agent safety and accountability.

By combining these three perspectives—**technical foundations, benchmarking practices, and governance theory**—this report aims to provide a **holistic analysis** of AI agents. The approach is deliberately interdisciplinary, bridging insights from computer science, economics, and law to capture both the opportunities and challenges of deploying agents in real-world contexts. This report adopts a synthesized review approach, combining general perspectives on AI agents with focused case studies of Auto-GPT and Devin AI.

# Findings

The synthesis of prior research reveals that AI agents occupy a unique position at the intersection of technical innovation, evaluation methodologies, and governance concerns. The findings can be grouped into three broad domains:

## Capabilities and Potential of AI Agents:

AI agents extend beyond traditional AI systems by embedding autonomy, flexibility, and adaptability. They are capable of setting goals, making decisions, and executing tasks in uncertain, dynamic environments without constant human supervision

. Unlike conventional object-oriented programs, agents can choose whether or not to execute a method based on internal states such as beliefs, desires, and intentions (the BDI model).

Recent advancements in LLM-driven agents have further expanded these capabilities. Agents can now plan multi-step actions, employ external tools, and interact through natural language

. This evolution positions agents not merely as assistants but as semi-autonomous collaborators, able to pursue open-ended goals such as conducting market research, debugging code, or booking complex travel itineraries

. The ability to adapt and generalize across domains suggests that agents will play a central role in future applications, ranging from commerce to governance.

## Capabilities:

From my study, I observed that Auto-GPT demonstrates how an open-source agent can autonomously break down goals into smaller tasks and make use of external tools like web browsing and file storage. This shows the **potential of agents to function without continuous human input**, which marks a clear shift from traditional chatbots or static AI models. On the other hand, Devin AI represents a more advanced, industry-focused application of agentic systems. It is capable of planning, writing, debugging, and deploying code in real-world environments. In my view, these two case studies show that agents are not only theoretical constructs but are already beginning to perform tasks of meaningful complexity.

When looking at how these agents are evaluated, I noticed a recurring problem. Auto-GPT gained wide attention due to its flexibility, but its **real performance is often inconsistent**, with issues like looping and hallucinations. Devin AI, while impressive in demonstrations, has struggled in independent tests, sometimes failing to complete tasks it was designed for. This confirms what I found in the literature: current benchmarks often emphasize accuracy or flashy demos, while neglecting cost, reproducibility, and long-term reliability. For me, this gap suggests that **evaluation methods must evolve** if we want to properly assess agents like Auto-GPT and Devin AI.

## Figures and Tables

|  |  |  |
| --- | --- | --- |
| **Feature** | **Auto-GPT** | **Devin AI** |
| **Core Function** | Open-source autonomous agent for general tasks | AI “software engineer” for end-to-end coding workflows |
| **+Autonomy** | Breaks down goals into sub-tasks, executes without constant supervision | Plans, writes, debugs, and deploys code independently |
| **Tool Use** | Web browsing, file management, memory, APIs | GitHub, VS Code, terminal, browsers, developer tools |
| **Applications** | Market research, content generation, business task automation | Bug fixing, app development, project deployment |
| **Significance** | First popular proof-of-concept for autonomous agents | First industrial-scale attempt at an AI-driven software engineer |

Table 1 illustrates how the literature on AI agents can be organized into three complementary perspectives: **state-of-the-art development, benchmarking practices, and governance frameworks**. The state-of-the-art perspective highlights the technical foundations of agents, such as autonomy, adaptability, and interaction in multi-agent environments. The benchmarking perspective sheds light on empirical evaluation, drawing attention to the limitations of current practices that emphasize accuracy while neglecting cost, reproducibility, and robustness. Finally, the governance perspective introduces legal and economic frameworks that address accountability, loyalty, and liability in the deployment of AI agents.

By juxtaposing these perspectives, several insights emerge. First, technological progress has outpaced both evaluation and governance mechanisms, resulting in a mismatch between agent capabilities and the standards used to assess or regulate them. Second, the need for **integration** is evident: effective progress in AI agents requires not only advances in algorithms but also robust benchmarks and governance structures. Third, the emerging consensus across domains emphasizes the importance of **multi-dimensional evaluation**—balancing technical performance with economic efficiency and ethical accountability.

This comparative overview underscores that the future of AI agents lies in **bridging the technical, empirical, and governance domains**, ensuring that systems are not only capable but also efficient, reliable, and trustworthy.

# Challenges

Despite significant advancements, the development and deployment of AI agents face several persistent challenges spanning technical, economic, and governance dimensions.

A. Technical Challenges

Early agent-oriented systems were hindered by the absence of standardized methodologies for design, verification, and implementation. Many systems were built in an ad hoc fashion, limiting their scalability and reusability

. While LLM-powered agents have addressed some of these issues, they introduce new challenges such as uncertainty in behavior and difficulties in ensuring consistent, predictable performance. Moreover, agents deployed in dynamic environments often encounter novel situations that cannot be fully anticipated during design, necessitating robust adaptability and continual learning.

B. Benchmarking and Evaluation Challenges

Benchmarking practices are central to measuring agent performance, yet current approaches suffer from critical shortcomings. Most prominently, there is an overemphasis on accuracy, which can be artificially inflated through repeated model queries or by exploiting benchmark-specific shortcuts

. This focus neglects cost-efficiency, which is crucial for real-world adoption, as many state-of-the-art agents are prohibitively expensive to operate at scale. Additionally, overfitting and reproducibility issues undermine the validity of benchmark results. The lack of standardized evaluation protocols makes it difficult to compare agents fairly, impeding meaningful progress.

C. Governance and Legal Challenges

Governance frameworks for AI agents are underdeveloped. The principal-agent problem, long studied in economics, highlights risks of information asymmetry, discretionary authority, and loyalty—all of which are amplified when agents act autonomously

. Traditional oversight mechanisms such as monitoring or incentive design may be insufficient for systems that operate at superhuman speed or execute unintuitive actions. Legal challenges further complicate the picture: it remains unclear who should be held liable when agents make harmful or unethical decisions. Without robust legal and regulatory frameworks, there is a danger of accountability gaps that could undermine trust in the technology.

D. Societal and Ethical Concerns

Beyond technical and legal domains, AI agents pose broader societal risks. Their ability to act independently raises concerns about labor displacement, privacy violations, and malicious use, such as automated cyberattacks

. Delegating economic and social activities to agents could reshape human behavior and social norms, creating systemic impacts that extend beyond individual users. Ethical concerns also arise regarding inclusivity—ensuring that agents serve not only their direct users but also respect broader societal values and rights.

E. Integration Gap

Perhaps the most significant challenge is the integration gap between these dimensions. While technical research advances rapidly, evaluation practices lag behind, and governance mechanisms remain underdeveloped. This fragmentation risks creating a future where powerful agents exist without adequate means of assessing their safety, efficiency, or ethical alignment.

# Conclusion

AI agents represent a significant shift in the evolution of artificial intelligence, moving beyond static, tool-like systems toward autonomous, adaptive, and collaborative entities. As this report has shown, research across three domains—technical development, benchmarking, and governance—offers complementary insights into both the potential and risks of this emerging paradigm.

From a technical perspective, agents demonstrate remarkable capabilities in autonomy, adaptability, and multi-agent interaction, particularly when powered by large language models. Yet, their deployment is hindered by the absence of standardized design methodologies and the unpredictability of behavior in dynamic environments

. Benchmarking studies highlight the limitations of current evaluation practices, which prioritize accuracy at the expense of cost, reproducibility, and generalizability, raising concerns about the practical utility of many state-of-the-art systems

. At the same time, governance and legal analyses emphasize unresolved issues of accountability, liability, and ethical alignment, stressing that traditional oversight mechanisms are inadequate for regulating autonomous agents operating at scale.

The overarching conclusion is that the promise of AI agents cannot be realized in isolation. Progress requires integration across technical, empirical, and governance dimensions. Future work must focus on cost-aware and reproducible benchmarking frameworks, robust agent design methodologies, and governance structures that ensure inclusivity, visibility, and liability. Only by bridging these domains can AI agents fulfill their potential as reliable, safe, and trustworthy collaborators in both personal and professional contexts.

In short, while AI agents are poised to become a cornerstone of future intelligent systems, their responsible advancement depends equally on innovation, evaluation, and regulation.

##### References

1. [1] E. Alonso, “AI and Agents: State of the Art,” AI Magazine, vol. 23, no. 3, pp. 25–32, Fall 2002.
2. [2] S. Kapoor, B. Stroebl, Z. S. Siegel, N. Nadgir, and A. Narayanan, “AI Agents That Matter,” Princeton University, Jul. 2, 2024. [Online]. Available: https://arxiv.org/abs/2407.01502
3. [3] N. Kolt, “Governing AI Agents,” Notre Dame Law Review (forthcoming), 2025.