California Management Review

ARTIFICIAL INTELLIGENCE

Rethinking AI Agents: A Principal-Agent Perspective

by Mohammad Hossein Jarrahi and Paavo Ritala

Image Credit | Sono Creative

Rethinking AI agents as guided actors: Balancing autonomy and accountability in organizations

☑ INSIGHT | FRONTIER 23 Jul 2025

The promise of AI agents has captured the business world's imagination, dominating headlines and strategic discussions across industries. In fact, much of the AI discussion during 2025 will be about AI agents. This begs the question: What is unique about AI agents and how will they transform your company?

RELATED CMR ARTICLES

Marcus Holgersson, Linus Dahlander, Henry W. Chesbrough, and Marcel Bogers, "**Open Innovation in the Age of AI**." California Management Review, 67/1 (2024): 5-20.

Jan Recker, Frederik von Briel, Youngjin Yoo, Varun Nagaraj, and Mickey McManus, "**Orchestrating human-machine designer ensembles during product innovation**." California Management Review 65/3 (2023): 27-47.

AI agents not only reason and learn but also interact directly with the environment, making decisions and executing them without direct human input. In other words, they **bridge knowledge and action** by proactively orchestrating complex workflows.

Today's AI agents also differ from agents of the past (with their well-structured inputs and outputs), as they can be much more adaptive across a wide range of granular tasks (where fixed paths cannot be hardcoded) based on foundation models. In short, the new AI agents should be able to deal with uncertainty, combine different tasks and workflows, coordinate across different systems and tools, and do all this within changing environments.

A simple example of an AI agent could be a personal AI agent that autonomously books flights, reserves hotels, and coordinates itineraries based on user preferences. More advanced examples of an AI agent in action could be onboarding clients, approving expenses, or customer service agents in retail. Unlike traditional chatbots, AI agents can autonomously interpret customer intent, process refunds, update shipping details, or escalate issues and exceptions to human supervisors—seamlessly combining reasoning, decision-making, and execution. In B2B settings, agentic AI systems can integrate data on industry trends with internal data on key customers' needs, therefore improving the ability to target and tailor offerings to corporate clients.

Companies like Google, OpenAI, and Anthropic have recently debuted increasingly sophisticated AI agents, showcasing how **the next competition ground** in foundation models has shifted to the terrain of AI agents. While Silicon Valley CEOs predict AI agents will outnumber **humans** soon and be used in **billions**, most organizations are still grappling with fundamental questions about what these agents are and how they should be deployed efficiently and responsibly. As such, the gap between the visionary rhetoric of autonomous AI agents and the practical realities of implementation remains wide. Despite rising interest and investment in AI agents and tech giants increasingly facilitating the infrastructure that enables **"open scalable agentic systems"**, organizations face difficult questions about their role, functions, and governance. Questions such as "**Can they actually be trusted** with something serious" will shape the efficiency and effectiveness gains available from agentic systems.

The concept of AI agents is still shrouded **in confusion**, with definitions varying widely across disciplines and applications. This definitional ambiguity creates challenges, especially in high-stakes organizational settings in which full autonomy is not always desirable or practical. Instead, it is more helpful to view AI agents as systems that integrate reasoning, decision-making, and execution *on behalf* of users. Specifically, we argue that looking at the role and function of AI agents through the lens of **principal-agent relationships** offers a more actionable perspective in business and organizational contexts.

The **principal-agent framework**, as traditionally defined in economic and organizational theory, describes a relationship where the principal delegates tasks to the agent, who acts on the principal's behalf[1]. Applying this lens to AI agents emphasizes intelligent agents working for humans and their role as systems designed to fulfill specific objectives while operating within human-defined constraints. From this perspective, AI agents function as intermediaries that balance autonomy with alignment to organizational goals.

Dimensions of AI Agents in Organizations

In what follows, we discuss key guiding principles for effective AI agent development and integration into business (see Figure 1), which can help organizations prepare for both efficient and responsible adoption while addressing key challenges.

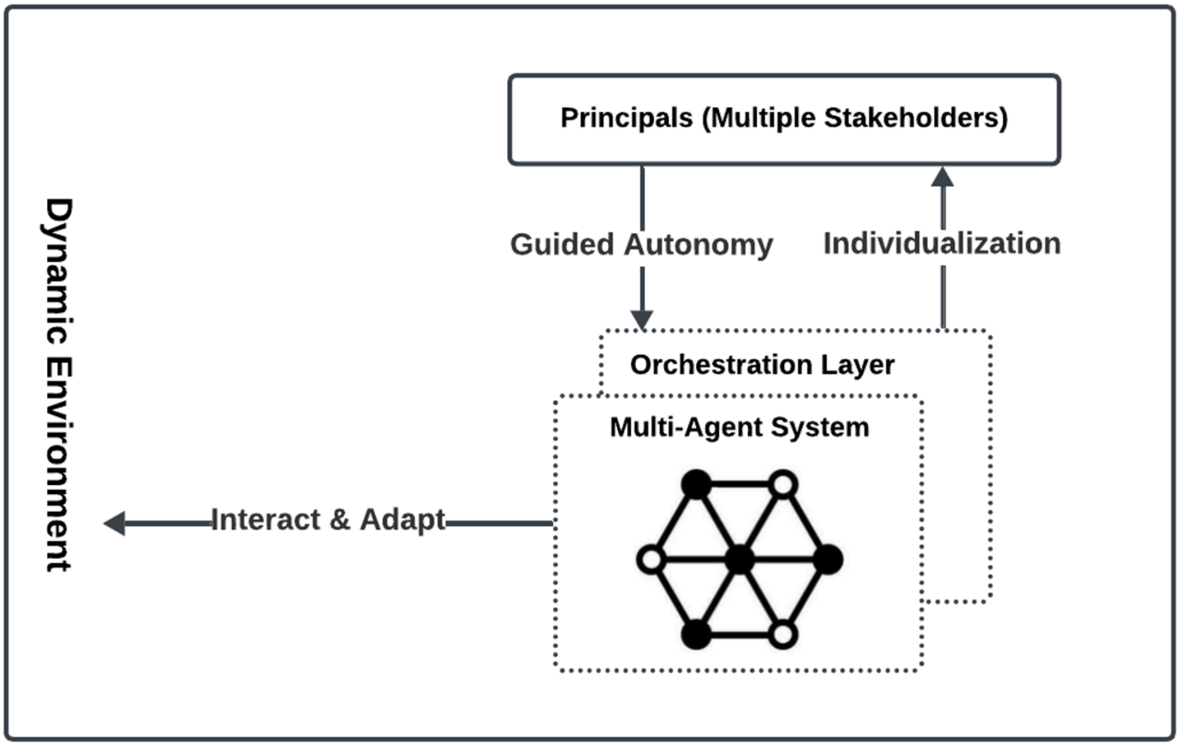


Figure 1: Key Dimensions of AI Agents from a Principal-Agent Perspective

Interactivity and Adaptability

Effective AI agents sense the environment, adapt to dynamic conditions, and proactively address challenges without waiting for explicit instructions. To be adaptive and transcend their training data, the agents must be given the latitude **to learn and grow**[2] in real-world situations. Because of their collaboration with other tools and systems, they can interact dynamically with the outside environment and act as active participants in organizational processes. This ability to integrate contextual cues and respond in real time improves their utility in changing environments. For example, a customer service AI can detect patterns of dissatisfaction in real time and escalate or even address issues before they become widespread. Furthermore, its ability to learn from interactions allows it to improve responses and anticipate common queries. Another example is the use of AI agents in cybersecurity, where the different agents are searching for anomalies in the process and customer data of the organization, and report back whenever potential suspicious activity is detected.

Guided Autonomy

While AI agents of the future are expected to achieve full autonomy, this is not always feasible or desirable in practice. Therefore, to act efficiently and responsibly, AI agents must strike a balance between autonomy and human oversight. Unlike fully autonomous systems, the principle of guided autonomy is about giving agents the leeway to execute decisions within defined boundaries of delegation. This approach provides a clear framework within which the agent operates, reducing risks associated with unrestrained autonomy when AI agents may face unknown unknowns. The boundaries of delegation can grow as agents expand their learning. This guided approach enables humans to continuously express their goals, monitor the agent's behavior[3], and effectively provide feedback. For example, in supply chain management, an AI agent may autonomously reallocate resources to address logistical bottlenecks and know when to escalate significant disruptions to human managers for resolution. In the domain of marketing and communications, AI agents could be developing targeted messaging based on customers' product use, while enabling human oversight on the process.

Collective Intelligence of Multiple Agents

There is increasing understanding that AI agents work better when they are specialized, rather than "multitasking" - in this sense, managing a multi-agent system resembles managing a multi-disciplinary team of professionals. When **working together** as a team or "swarm", AI agents must interact with other agents, enterprise systems, processes, and stakeholders, while relying on human guidance where necessary, to function effectively within the volatile organizational environment. For example, generative AI systems (capable of generating new content) and analytical AI systems[4] (with classification and prediction capabilities) can leverage the strengths of different system architectures to automate workflows. For example, in financial institutions, AI agents managing risk assessments communicate with other agents which handle compliance to ensure that investments adhere to regulatory requirements while maximizing returns. In the domain of market research, there might be a separate “research agent” and “reporting agent”, whose workflows are combined to provide integrated reports to different audiences and domain experts inside the company.

Safety, Accountability, and Interoperability through Orchestration

Similarly as in human principal-agent relationships, AI systems require mechanisms to ensure their actions are safe, predictable, and accountable. However, generative AI foundation models often exhibit surprising performances that can be unpredictable, inconsistent, and even erratic. While they may be useful for simple tasks, these behaviors can prove catastrophic for more critical organizational processes, particularly when autonomous execution is concerned. A useful strategy is to combine these models with rule-based control mechanisms and structured reasoning[5] to mitigate the erratic behavior of generative AI systems, which may include not only hallucinations[6] but even deceptive actions[7]. As AI pioneer **Yoshua Bengio** recently noted, “The good news is that if we build non-agentic systems, they can be used to control agentic systems.”

Another important strategy in the orchestration layer is standardization, including shared protocols and data formats, to ensure compatibility and communication between diverse agents and systems. This includes a technical layer that integrates with existing technology stacks and communication with other AI agents but also shared protocols located outside of AI agents[8] that guide how agents may interact with institutions like legal systems.

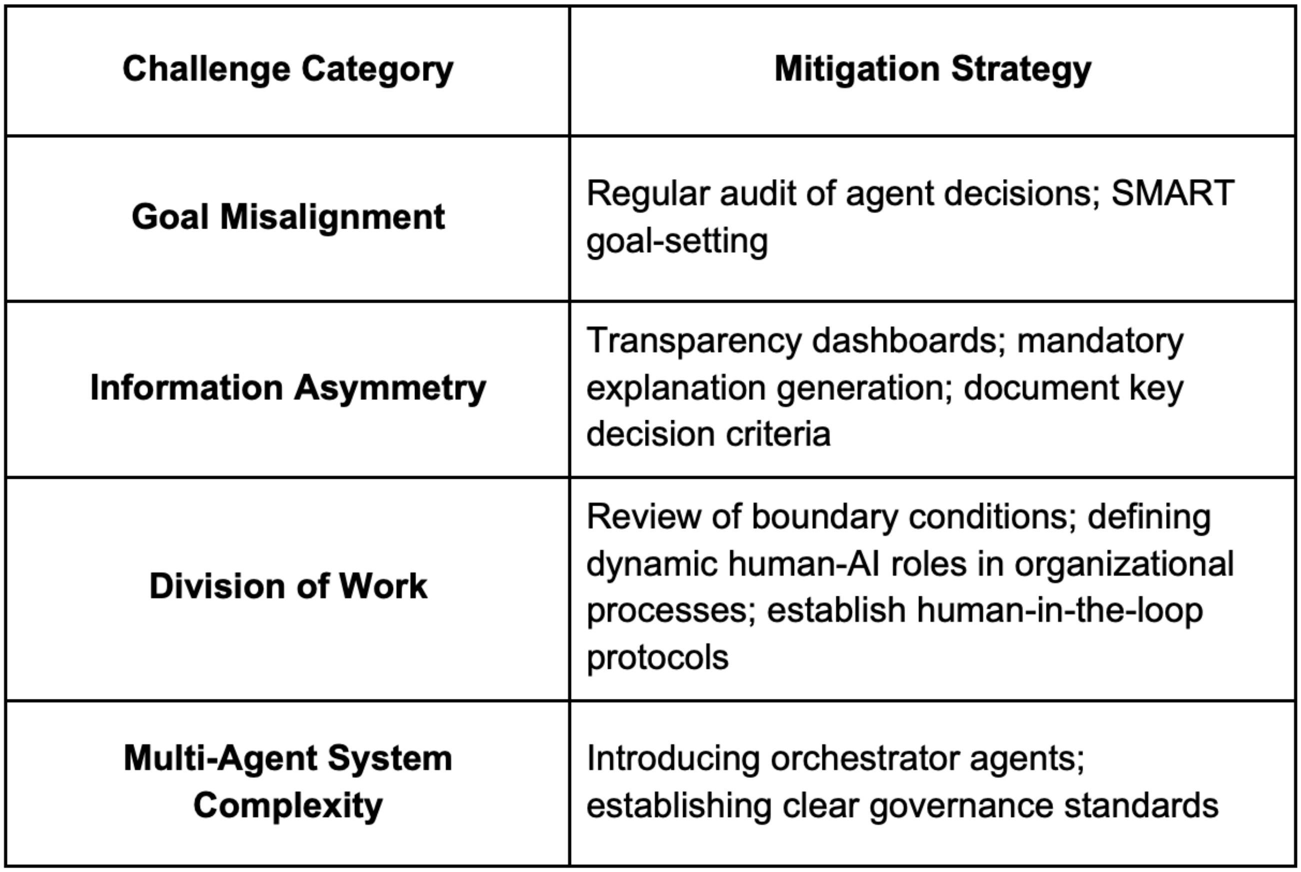
Individualization and Alignment

AI agents must tailor their actions to the needs and goals of specific stakeholders - a phenomenon known as the **alignment problem**. Getting AI alignment right requires an “individualization engine,” where AI agents can cater to the needs and objectives of stakeholders and dynamically adjust their outputs to reflect various individual priorities and organizational goals (e.g., profit maximization versus environmental sustainability). As an example, AI agents in sales can assist sales representatives by learning their preferred negotiation tactics and customer engagement strategies, as well as creating personalized campaigns by analyzing different customers' data and aligning their recommendations accordingly.

Challenges in Principal-Agent Dynamics

Despite their potential, AI agents introduce unique challenges that mirror many traditional principal-agent problems. To address these systematically, organizations should implement risk mitigation strategies (see Table 1).

Table 1: Challenges in Principal-Agent Dynamics and relevant mitigation strategies



Challenge Category	**Mitigation Strategy**
Goal Misalignment	Regular audit of agent decisions; SMART goal-setting
Information Asymmetry	Transparency dashboards; mandatory explanation generation; document key decision criteria
Division of Work	Review of boundary conditions; defining dynamic human-AI roles in organizational processes; establish human-in-the-loop protocols
Multi-Agent System Complexity	Introducing orchestrator agents; establishing clear governance standards

Goal Misalignment

The goals carried out by the AI agent may diverge from the broader strategic objectives of the organization, particularly if incentives or constraints are misaligned. For example, an AI agent optimizing sales might prioritize upselling products, inadvertently damaging customer trust or satisfaction.

The risk mitigation strategy may involve defining the goals of the agents through approaches such as the SMART approach[9] (specific, measurable, achievable, relevant, time-bound) and implementing regular audits of agent decisions. For instance, a retail AI agent should have clear upper limits on discount offerings and must maintain a minimum customer satisfaction score while optimizing for sales. Furthermore, if there is no continuous human-in-the-loop strategy, routine or ad-hoc audits of agentic workflows and decisions should be conducted.

Information Asymmetry

AI agents often have access to vast amounts of data and information processing capabilities that their human principals cannot fully interpret or monitor. This can lead to potential misuse or “overinterpretation” of data. For example, a hiring AI might leverage hidden correlations in data to make decisions, potentially perpetuating biases that may go unnoticed or unchecked by the organization.

The risk mitigation strategy could be to deploy transparency dashboards[10] that track key decision variables and require the AI to generate plain-language explanations for all critical decisions. For hiring scenarios, this might include mandatory documentation of the specific qualifications and criteria used for each candidate’s evaluation and may involve the application of another agent as an explainability agent.

Division of Work

One of the key challenges in applying agentic AI is defining the dynamic division of labor between humans and AI agents and finding the right places for AI augmentation and automation[11]. For instance, what types of problems require human input? Are humans only involved in handling exceptions, or do they play a broader role? The confusion about responsibility in automated systems can result in a phenomenon described as a "moral crumple zone[12]", where responsibility is diffused and misinterpreted between humans and agents.

The risk mitigation strategy involves carefully articulating the boundary conditions under which decision-making can be reliably delegated to agentic AI systems and the circumstances in which human decision-makers must get involved (i.e., human-in-the-loop). Effective business process management[13] can be particularly useful for developing the right combination of human resources and agentic roles, much like teammates in specific organizational processes, yet considering also the inherent differences in human and AI capabilities.

Multi-Agent System Complexity

It is gradually becoming clear that AI agents work best when specialized agents are combined into a multi-agent system. However, the complexity of interactions between multiple agents can result in new unpredictabilities, particularly when they reflect competing goals and may use different technical architectures. Too many agents with too many tasks can result in overly complex and difficult-to understand systems. For example, Moderna recently deployed 3,000 '**tailored GPTs**' to support specific tasks (e.g. dose selection for clinical trials and drafting responses to regulatory questions). While these tools can effectively assist with individual tasks (the trees), they risk losing sight of the broader strategic context (the forest). These challenges require new approaches to multi-agent governance with clear standards for accountability and transparency. Beyond standardizing the multi-agent environment, there is a need for **orchestrator agents**, which serve as the entity that coordinates the interactions of multiple AI agents, improving overall efficiency and effectiveness. Accenture calls these **"strategic agents**," which act as a team leader by coordinating multiple utility agents. As an example, in a logistics network, an orchestrator agent can ensure multiple agents (in charge of handling inventory, routing, etc.) communicate effectively and align their tasks to optimize the whole supply chain.

In addition, as multi-agent systems gradually take on more complex tasks and important responsibilities, an organization-wide AI governance strategy is needed to build guardrails for agents to maintain coherence and safety. For example, **at McKinsey**, a central team reviews all developed agents based on risk, legal, and data policies before they are rolled out.

Conclusions

AI agents bring us one step closer to the idealist vision of AI as **a true partner, not just a tool**. However, adopting AI agents requires an AI governance regime that goes beyond the technical capabilities of these systems, integrates them into organizational processes and ensures adaptability, interoperability, and the safety of AI agents as collective swarm systems.

By adopting the principal-agent perspective, companies and their leaders can create more accountable, effective, and responsive AI agents. Ultimately, AI agents become embedded in all the core workflows of organizations and will be augmenting most if not all human work. The principles such as guided autonomy, individualization, and adaptability will help design such organizations.

References

1. Caers, R., Bois, C. D., Jegers, M., Gieter, S. D., Schepers, C., & Pepermans, R. “Principal-Agent Relationships on the Stewardship-Agency Axis.” Nonprofit Management and Leadership, 17/1 (2006): 25-47.
2. Walsh, M. “How Much Supervision Should Companies Give AI Agents?”Harvard Business Review, (2025).

3. Bansal, G., Vaughan, J. W., Amershi, S., Horvitz, E., Fourney, A., Mozannar, H., & Weld, D. S. "Challenges in Human-Agent Communication." arXiv preprint arXiv:2412.10380, (2024).
4. Davenport, T. and High, P. "How Gen AI and Analytical AI Differ — and When to Use Each. Harvard Business Review, (2024).
5. Shah, C., & White, R. W. "Agents are not enough." arXiv preprint arXiv:2412.16241, (2024).
6. Hannigan, T. R., McCarthy, I. P., & Spicer, A. "Beware of botshit: How to manage the epistemic risks of generative chatbots." Business Horizons, 67/5 (2024): 471-486.
7. Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., & Hobbhahn, M. "Frontier models are capable of in-context scheming." arXiv preprint arXiv:2412.04984, (2024).
8. Chan, A., Wei, K., Huang, S., Rajkumar, N., Perrier, E., Lazar, S., ... & Anderljung, M. "Infrastructure for AI Agents." arXiv preprint arXiv:2501.10114, (2025).
9. Purdy, M. "What Is Agentic AI, and How Will It Change Work?" Harvard Business Review, (2024).
10. Chen, Y., Wu, A., DePodesta, T., Yeh, C., Li, K., Marin, N. C., ... & Viégas, F. "Designing a dashboard for transparency and control of conversational AI." arXiv preprint arXiv:2406.07882, (2024).
11. Ritala, P., Ruokonen, M., & Ramaul, L. "Transforming boundaries: how does ChatGPT change knowledge work?" Journal of Business Strategy, 45/3 (2023): 214-220.
12. Elish, M. C. "Moral crumple zones: Cautionary tales in human-robot interaction." Engaging Science, Technology, and Society, 5/0 (2019): 40-60.
13. Davenport, T. and Redman, P. "How to Marry Process Management and AI." Harvard Business Review, (2025).

Mohammad Hossein Jarrahi Follow

Mohammad Hossein Jarrahi is a Professor of Information Science at UNC Chapel Hill. His research focuses on the intersection of work, business, and AI, with a particular emphasis on human-AI symbiosis in the future of work and management.

Paavo Ritala Follow

Paavo Ritala is a Professor of Strategy and Innovation at the Business School at LUT University, Finland. His main research themes include ecosystems and platforms, the role of data, algorithms, and digital technologies in organizations, and circular and regenerative economy.