



Shaping the future of multi-agent systems with responsible AI



The better the question. The better the answer.
The better the world works.

**EY**

Shape the future
with confidence

Contents

- Executive summary2
- Introduction: multi-agent systems3
- Use cases highlighting responsible AI challenges with MAS.....5
- Why responsible AI is difficult to achieve in MAS6
- How to implement a future-focused plan for MAS8

Executive summary

- As artificial intelligence rapidly advances, autonomous agents are beginning to collaborate and operate in multi-agent systems (MAS), which introduce unique challenges with respect to responsible AI.
- For instance, different agents in customer service all may be focused on unique tasks that could conflict or rely on context that is lost, and hallucinations can compound, leading to confusion and dissatisfaction.
- Frameworks should be introduced to manage the implications of MAS effectively. In these early days of the technology, human oversight and accountability are among the actions we recommend for responsible AI implementations.



Introduction: multi-agent systems

In just the past few years, artificial intelligence (AI) has been revolutionized. Sophisticated language models are enabling use cases such as chatbots and automated content generation, putting the technology in the hands of everyday people through natural conversational capabilities that significantly improve user engagement.

Now the buzz centers on autonomous agents, which perform specific tasks independently, making decisions based on their programming and incoming data – and even work with other agents. But the complexity raises the stakes for making these systems safe and ethical.

In multi-agent systems (MAS), agents collaborate, negotiate and compete to achieve individual or collective goals. This paradigm shift opens new possibilities for automation and optimization across various domains, including robotics, finance, transportation and health care.

The effectiveness of agents is shaped by five core factors that define how they operate, interact and evolve. Understanding these dimensions is essential for designing, deploying and governing AI responsibly, as each factor carries specific implications for safety, trust and performance.

- **Autonomy** captures the degree to which an agent can act and make decisions independently. Highly autonomous agents can sense their environment, evaluate options and act with minimal human intervention, driving efficiency and responsiveness at scale. However, this independence heightens the need for robust oversight, ethical safeguards and accountability frameworks. At the other end of the spectrum, low-autonomy agents are capable of executing basic, repetitive tasks but depend heavily on humans for guidance in novel or complex situations.
- *High autonomy*: Minimal human involvement, enabling scalability but requiring strong governance.
- *Low autonomy*: Limited independence, relying on frequent human direction.

(cont.)

- **Adaptability** reflects an agent's ability to adjust its decision-making and behavior in response to changing circumstances. Adaptive agents can learn from historical data, user interactions and environmental cues, continuously refining their strategies to remain effective in dynamic contexts. This flexibility enables resilience but also introduces challenges in monitoring for unintended bias or drift. In contrast, rigid agents operate within a narrow, predefined rule set, making them predictable but potentially brittle when faced with unexpected scenarios.

- *Adaptive*: Continuously learns and adapts to new inputs and conditions.

- *Rigid*: Operates with fixed logic and limited capacity to evolve.

- **Collaboration capability** defines an agent's ability to operate effectively within a multi-agent or human-machine ecosystem. Collaborative agents are designed to exchange information, negotiate and coordinate their actions, making them essential for complex workflows such as supply chain optimization, real-time trading systems or customer service orchestration. Independent agents, by contrast, function as standalone systems, suitable for contexts where interaction is unnecessary or adds complexity.

- *Collaborative*: Works seamlessly with humans or other agents to achieve shared goals.

- *Independent*: Operates in isolation with minimal or no need for coordination.

- **Temporal stability** refers to both the expected lifespan of an agent and its consistency over time. Persistent agents are designed to operate indefinitely, maintaining a stable identity and performance profile. This consistency is critical for trust in mission-critical applications such as health care, finance and defense. Transient agents, on the other hand, are temporary by design, created to fulfill a single objective or function for a limited duration. This makes them ideal for short-lived scenarios like troubleshooting events, ad-hoc data analysis or time-bound customer interactions.

- *Persistent*: Long-term operation with stable behavior, enabling reliability and predictability.

- *Transient*: Purpose-built for temporary, time-bound or goal-specific tasks.


- **Reentrancy** measures an agent's ability to handle interruptions and resume progress seamlessly. Reentrant agents are designed for dynamic, high-stakes environments where interruptions, such as external alerts, user overrides or system events, are common. These agents can pause, retain context and resume operations without loss of functionality, which is vital in fields like logistics, emergency response and algorithmic trading. Non-reentrant agents, however, require uninterrupted execution; disruptions force a restart or failure, making them simpler but less flexible.

- *Reentrant*: Supports interruptions while maintaining task continuity.

- *Non-reentrant*: Cannot resume tasks if interrupted, requiring a full restart.

Together, these characteristics contribute to the creation of robust and reliable agents that can thrive in complex, dynamic environments – but the decentralized nature of MAS complicates traditional accountability frameworks, as responsibility is often distributed among multiple agents. Furthermore, the potential for conflicting objectives among agents can lead to ethical dilemmas that require careful navigation. Recognizing and addressing these challenges is crucial for ensuring multi-agent systems operate safely and responsibly in real-world environments.





Use cases highlighting responsible AI challenges with MAS

What do we mean by MAS in practice — and what potential pitfalls can arise from their interactions? The following examples illustrate how emergent behaviors, coordination challenges and transparency issues can impact user experience and trust in AI systems.

Example 1

In the realm of customer service, chatbots serve as central interfaces, supported by multiple specialized agents responsible for different workflows, such as technical support and billing inquiries. For instance, when a customer reports a product issue, the chatbot engages a technical support agent for troubleshooting. If account details are needed, the inquiry is seamlessly transitioned to a billing agent who can verify subscription status. However, challenges can arise if these agents do not communicate effectively. If the technical support agent provides a solution that requires a specific software version, but the billing agent is unaware of this requirement, the customer may receive conflicting information, leading to confusion and dissatisfaction.

Example 2

Consider a system where multiple agents interact with users through dialogue. If one agent interprets a user's request differently than another, it can lead to a frustrating user experience. For example, if a user asks for restaurant recommendations and one agent suggests fast food while another recommends fine dining, the inconsistency can erode user trust. This situation highlights the critical importance of establishing clear communication protocols and value alignment among agents to ensure seamless and coherent interactions.

Example 3

If agents are designed to provide responses based on user interactions but lack a cohesive understanding of context across different exchanges, it can lead to inconsistencies. For instance, if a user asks a follow-up question that builds on a previous conversation, but the agents do not retain the context from earlier exchanges, the response may seem irrelevant or confusing. This inconsistency can frustrate users and diminish their trust in the system. And resolving these problems can be more difficult than you think.

A photograph of two men in a dark office environment. One man, wearing glasses and a blue blazer, is pointing at a tablet held by the other man, who is also wearing glasses and a dark t-shirt. They are both looking intently at the device. In the background, there are blurred lights and office equipment.

Why responsible AI is difficult to achieve in MAS

Because these systems are so complex and dynamic, organizations must comprehensively understand their potential risks and ethical implications before they are deployed. Key concerns — encompassing accountability, ethical alignment, performance, transparency and security — should be identified and carefully considered so that MAS are implemented successfully and operate effectively and responsibly.

Executives must view these systems through several lenses to implement and accelerate responsible AI protections:

- **Emergent behavior and accountability:** Interactions among autonomous agents can lead to unpredictable emergent behaviors, which are inherent characteristics of MAS. While these behaviors are a natural outcome of agent interactions, they raise significant accountability questions, especially when they result in harm or unintended consequences. This unpredictability complicates how clear lines of responsibility are established – who is liable when such situations arise? Addressing these accountability challenges is particularly critical in regulated environments like financial services, where the implications of emergent behavior can have far-reaching effects.
- **Coordination ethics and value alignment:** Conflicts may emerge between the individual goals of agents and the collective good within MAS. In the context of regulatory frameworks, agents should operate under aligned ethical standards and values. This alignment is necessary to maintain trust and compliance with regulatory expectations, particularly in sectors like finance.
- **Performance and cost:** The reliance on third-party models in MAS can lead to performance bottlenecks and increased operational costs, especially during peak demand periods. This reliance can affect the overall efficiency and sustainability of the system, making it vital for organizations to evaluate the cost-benefit ratio of integrating external models into their MAS.

(cont.)

- **Transparency and explainability of interactions:** The complexity of interactions among agents can create opaque decision-making processes, making it challenging for stakeholders to understand and audit these decisions. In regulated industries, this lack of transparency can undermine trust and compliance, requiring mechanisms for clearer communication and explainability of agent interactions.
- **Decentralized responsibility:** The distribution of responsibility among multiple agents complicates accountability when harm occurs. Establishing clear frameworks for responsibility is essential to navigate the complexities of decentralized systems, ensuring that all agents understand their roles and obligations, particularly in compliance-heavy environments.
- **Access control and secure response generation:** Ensuring appropriate access control across multiple agents is a key challenge because users may have permissions for some, but not all systems involved in a query. Responsible AI requires the system to handle such cases transparently and securely by deciding whether to provide partial answers, issue access errors, or offer explanatory fallback responses while balancing user trust, privacy, and fairness and preventing inadvertent disclosure of restricted information.
- **Autonomy vs. human control:** The limited centralized oversight in MAS raises questions about the extent of human control over autonomous agents. This lack of oversight can lead to ethical dilemmas and unintended consequences, making it essential for organizations to establish governance frameworks that balance autonomy with necessary human intervention.
- **Hallucinations:** Agents may generate misleading information, leading to inconsistencies and eroding trust, particularly when multiple agents rely on the same information sources. Addressing the risk of hallucinations is critical to maintaining the reliability of the system, especially in contexts where accurate information is essential for decision-making.





How to implement a future-focused plan for MAS

As multi-agent systems continue to evolve, it becomes paramount to address the unique implications of responsible AI so that the systems are deployed ethically, effectively and proactively — so that they function well but also uphold the values and principles that society holds dear.

Yet, to some extent, that reveals a tension between the autonomous agents of the future and the human oversight necessary to ground actions in our values and ethics. Today, it is crucial to balance human oversight with agent autonomy. While MAS can function with limited centralized control, mechanisms for effective human intervention must be in place to prioritize ethical considerations.

Other tactics that should be followed will differ from system to system, but some general ones apply:

- Establish clear frameworks for accountability, as they help stakeholders understand who is responsible for agents' actions, fostering trust and reliability. Clear accountability for each agent's role ensures accurate and timely responses, improving overall satisfaction and trust.
- Align the diverse ethical values of agents with the collective good, which will mitigate conflicts and promote cooperation, leading to more beneficial outcomes for society.
- Promote transparency in decision-making processes to enhance stakeholder understanding and enabling effective auditing of MAS. Coupled with robust security measures, these systems can protect against risks such as adversarial exploitation and data breaches, ensuring the integrity of both users and the systems themselves.

- Implement a shared communication protocol among agents, to mitigate against the problems shown in our illustrative use cases. Such protocols allow agents to update each other on the status of inquiries and share relevant information in real time.
- Utilize a centralized knowledge base where all agents can access up-to-date information about products, services and customer interactions to further enhance coordination.
- Implement a shared context management approach among agents so that they can maintain and update context in real time, ensuring that all parts of the system are aware of the ongoing conversation. Clear accountability for context retention and response generation further enhances the user experience, ensuring coherent and contextually appropriate responses.
- Acknowledge the uncertainties surrounding agent behaviors and the limitations of our current understanding. Continuous cross-disciplinary academic research on these responsible AI challenges related to MAS is essential for pushing forward our knowledge and improving system design.

By integrating principles of accountability, value alignment, transparency, security, and fairness, stakeholders can collaboratively develop MAS that operate efficiently and align with societal values. This commitment to responsible AI will ultimately enhance human well-being and foster a more equitable society.

Author



Mary Kryska

EY Americas AI and Data
Responsible AI Leader

mary.kryska@ey.com

Thank you to **Anil Sood** for contributing to the development of this article.

EY | Building a better working world

EY is building a better working world by creating new value for clients, people, society and the planet, while building trust in capital markets.

Enabled by data, AI and advanced technology, EY teams help clients shape the future with confidence and develop answers for the most pressing issues of today and tomorrow.

EY teams work across a full spectrum of services in assurance, consulting, tax, strategy and transactions. Fueled by sector insights, a globally connected, multidisciplinary network and diverse ecosystem partners, EY teams can provide services in more than 150 countries and territories.

All in to shape the future with confidence.

EY refers to the global organization, and may refer to one or more, of the member firms of Ernst & Young Global Limited, each of which is a separate legal entity. Ernst & Young Global Limited, a UK company limited by guarantee, does not provide services to clients. Information about how EY collects and uses personal data and a description of the rights individuals have under data protection legislation are available via ey.com/privacy. EY member firms do not practice law where prohibited by local laws. For more information about our organization, please visit ey.com.

Ernst & Young LLP is a client-serving member firm of Ernst & Young Global Limited operating in the US.

© 2025 Ernst & Young LLP.
All Rights Reserved.

US SCORE no. 28471-251US_2
BSC no. 2507-11008-CS
ED None

This material has been prepared for general informational purposes only and is not intended to be relied upon as accounting, tax, legal or other professional advice. Please refer to your advisors for specific advice.

ey.com