AI agent governance: Big challenges, big opportunities

Advanced AI agents don't just think—they *do*. Whereas previous generative AI (genAI) tools created content, made predictions or provided insights in response to human prompting, agents can go out into the world and accomplish complex tasks autonomously. Moreover, agents can make decisions on the fly and adapt to changing conditions. This presents new challenges for AI governance.

Artificial intelligence governance refers to the processes, standards and guardrails that help ensure AI systems and tools are safe and ethical. AI governance frameworks direct AI research, development and application to help ensure safety, fairness and respect for human rights.

When it comes to agents, governance frameworks will need to be updated to take the autonomy of agents into account. The economic potential for agents is vast, but so is the associated risk landscape. Encouraging intelligent systems to operate more safely, ethically and transparently will be a growing concern as they become more autonomous.

Industry newsletter

The latest AI trends, brought to you by experts

Get curated insights on the most important—and intriguing—AI news. Subscribe to our weekly Think newsletter. See the IBM Privacy Statement.
We use your email to validate you are who you say you are, to create your IBMid, and to contact you for account related matters.
Business email

Your subscription will be delivered in English. You will find an unsubscribe link in every newsletter. You can manage your subscriptions or unsubscribe here. Refer to our IBM Privacy Statement for more information.

Your subscription will be delivered in English. You will find an unsubscribe link in every newsletter. You can manage your subscriptions or unsubscribe here. Refer to our IBM Privacy Statement for more information.

Autonomous decision-making without human oversight

The very characteristics that make agentic AI powerful—autonomy, adaptability and complexity—also make agents more difficult to govern. One of the primary governance challenges of AI agents is their ability to make decisions independently. Unlike conventional software systems that follow strict, rule-based programming, AI agents use machine learning algorithms to analyze data and determine actions based on probabilities. This autonomy allows AI to operate in real-time environments.

This lack of human control makes it harder to ensure that AI agents act in a safe, fair and ethical way. In high-risk situations such as autonomous vehicles or algorithmic stock trading, an AI agent's decision can have major consequences, yet human oversight is not always available. This creates a governance

dilemma. How can leaders balance AI's efficiency and autonomy with the need for accountability and control?

Many AI agents, especially more advanced agents powered by machine learning, perform decision-making processes that aren't easy for humans to interpret. Unlike rule-based systems with traceable logic, machine learning models make decisions based on complex patterns in data that even their developers can't fully understand. This opacity makes it hard to audit AI-driven decisions, which is a challenge in fast-moving automation use cases. Imagine if an AI system were to deny a loan application based on bad data, or a healthcare system to recommend the wrong treatment. Stakeholders must be able to understand the rationale behind the decision.

Bias is another challenge. AI systems learn from historical data, but if the data contains biases, AI may amplify them. AI agents may make undesirable decisions such as prioritizing efficiency over fairness or privacy.

Security and compliance risks

Like any AI system, autonomous agents are also vulnerable to security threats. AI models and bots can be manipulated through adversarial attacks, where slight modifications to input data trick the AI into making incorrect decisions. Large language models (LLMs) and chatbots that communicate with users in natural language can be tricked into generating harmful content. The decentralized deployment of AI agents makes it difficult to implement uniform security measures.

Agentic systems often rely on APIs to integrate with external applications and data sources. Poorly governed APIs can expose vulnerabilities, making them targets for cyberattacks. Cybersecurity risks include adversarial attacks, data leaks and unauthorized access that exposes sensitive information. To mitigate these risks, APIs should have access controls and authentication mechanisms to prevent unauthorized interactions.

Apart from security, organizations also need to adhere to regulations when designing AI agents. However, regulations often lag behind technological advancements. AI systems are inherently complex and unpredictable, and compliance requirements can be ambiguous or contradictory. We may soon see the world's national and transnational governing bodies build rules around the use of agents specifically.

Navigating uncharted waters

Traditional AI governance best practices like data governance, risk assessments, transparent workflows, explainability, ethical standards and continuous monitoring also apply to agentic systems. But agentic governance can go beyond these established practices.

Instead of just testing models before deployment, organizations can create simulated environments where AI agents can make decisions without real-world consequences before being fully deployed. AI sandboxing allows developers to study unintended ethical dilemmas before exposing agents to real users. Ethical AI models can be tested under moral stress tests, such as simulated self-driving accident scenarios or ethical dilemmas in hiring AI.

Agent-to-agent monitoring is another way to head problems off before they get out of control. Because agentic ecosystems can be so complex, agents will need to collaborate and negotiate with one another often. Monitoring these interactions and establishing conflict resolution rules for agents can help ensure that they can work together in harmony.

Working agents can also be paired with "governance agents" designed to monitor and evaluate other agents, and prevent potential harm. For risk mitigation, agents must be continuously monitored to detect model drift. Imagine a customer service agent that deals with grumpy customers all day developing a bad-tempered personality as a result of adapting across such interactions. Now imagine a governance agent behaving like a hall monitor, pulling this agent aside and communicating something along the lines of, "You don't seem yourself today." Agents can also be programmed to seek human approval for certain actions.

Beyond these practices, many experts recommend that agents have an emergency shutdown mechanism that would allow them to be immediately deactivated, especially in high-risk environments. Organizations can establish containment procedures to help ensure that malfunctioning AI cannot escalate issues before intervention. Some organizations are experimenting with stress testing agents with adversarial attacks in edge cases and under extreme or unexpected conditions to identify vulnerabilities.

Governing AI agents will soon be a bit easier. Governance platform providers will offer robust AI governance tools with dashboards that provide access to specialized metrics for agentic systems and agent interaction. For example, software engineers at IBM are currently working on integrating specialized metrics such as context relevance, faithfulness and answer similarity into watsonx.gov. The right governance software will help stakeholders keep track of their agents across their end-to-end lifecycle, allowing them to get the most out of agentic AI.

As agentic AI systems become more autonomous, ensuring they operate safely, ethically and securely is a growing challenge. Organizations must adopt scalable governance models, enforce strong cybersecurity and risk management protocols and integrate human-in-the-loop oversight. If organizations can scale agentic systems safely, they'll be able to capture virtually limitless value.