VIEW POINT

GUARDRAILS IN AGENTIC AI SYSTEMS: DESIGNING GUARDRAILS AROUND AGENTIC AI

Abstract

Agentic AI systems can make autonomous decisions and work toward complex goals. They represent a revolutionary shift that will transform how businesses operate. However, AI agents pose significant risks including hallucinations in critical systems, misaligned long-term goals, and potential competition with humans for resources. Such risks make robust safety measures essential. Companies must implement comprehensive guardrails including governance frameworks, multi-layered technical controls, and operational safeguards to safely deploy these powerful autonomous AI systems, while maintaining human oversight.

Infosys®
Navigate your next

This technological advance feels different! As enterprises stand at the precipice of the next major technological shift, many veterans of the technology industry from Argenti to NVIDIA's Founder Jensen Huang have commented that what the world is experiencing with Artificial Intelligence (AI) is not an evolution, but a revolution. Agentic AI is an incremental revolution within this shift that represents both unprecedented opportunity and formidable risk. The same Marco Argenti recently noted that AI capabilities to plan and execute complex, long-running tasks on humans' behalf will begin to mature in 2025, creating conditions for companies to eventually "employ" and train AI workers as part of hybrid teams.

Promising? Definitely. Ominous? Terribly! Any inexperienced technology worker will wonder at the implications of working alongside robotic co-workers. The implications are profound for safety, ethics as well as resilience in settings ranging from typical back office work to hazmat situations, medtech environments, or air traffic control. As agentic AI marches into workplaces, this is the critical imperative: the establishment of robust guardrails to ensure safe, ethical, and resilient deployment.

Let's start at the beginning. Agentic AI differs fundamentally from other generative AI tools such as Large Language Models (LLMs) that have become part of the discourse today. Agentic AI systems are autonomous systems where multiple independent AI agents, each capable of making decisions, work to achieve complex goals through collaboration, coordination, or even competition with other systems.

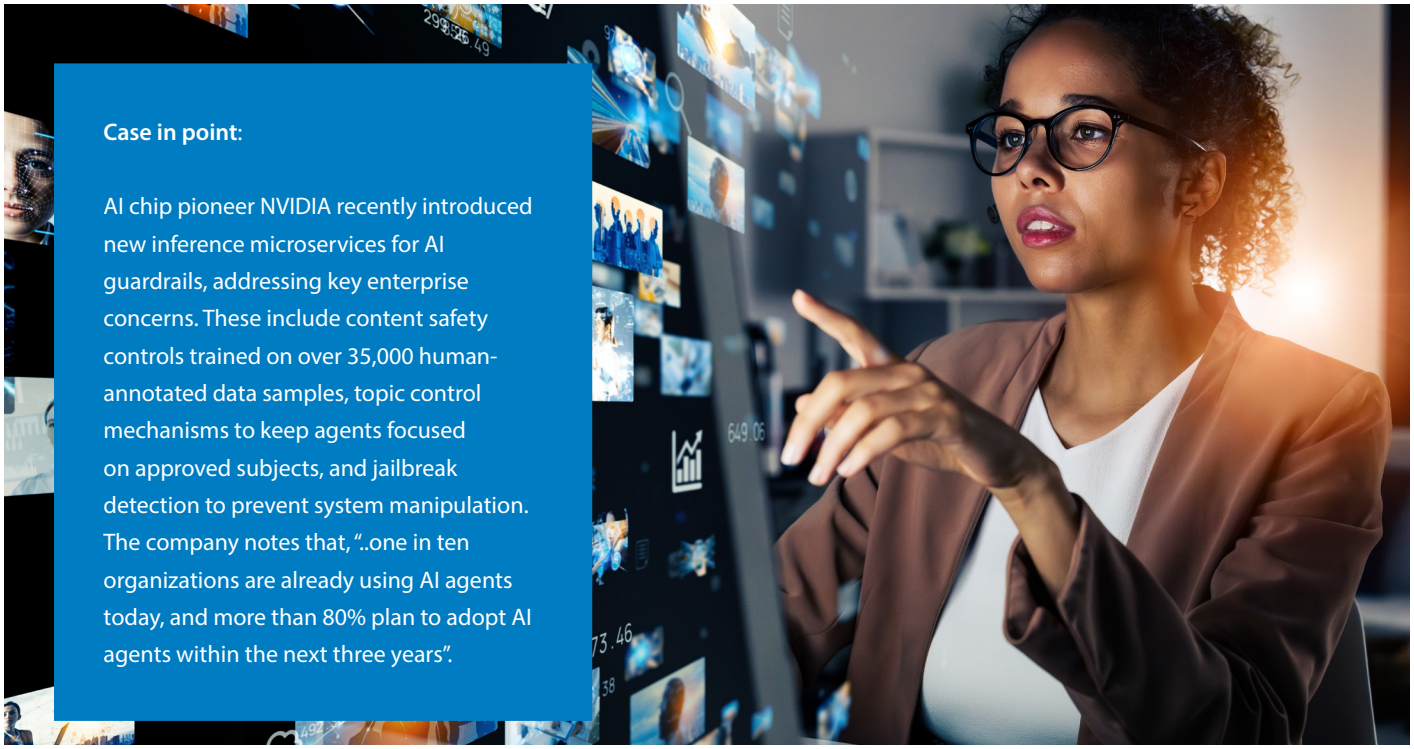
Quite unlike traditional chatbots that respond to prompts, agentic AI systems are autonomous entities that perceive their environment, make decisions, and manifest "agency" to achieve specific goals. Then, we need to understand multi-agent AI, the next shift-within-a-shift. These are networks of multiple autonomous agents that can interact, coordinate, and collaborate to achieve individual or collective goals. Microsoft's Ray Smith, the company's VP of AI Agents, emphasizes that multi-agent invocation and debugging capabilities are essential for enterprise needs. He noted recently that, "..it's very hard to create a reliable process that you squeeze into one agent. Breaking it up into parts improves maintainability and makes building solutions easier, but also significantly enhances reliability".

The enterprise impact of agentic AI is projected to be substantial. Deloitte predicts that 25% of enterprises using generative AI will deploy AI agents in 2025, growing to 50% by 2027. IDC's recent survey found that more than 80% of companies believe "AI agents are the new enterprise apps, triggering reconsidering of investments in packaged apps".

Leading AI researchers have warned about the risks with different types of AI. We know LLMs are prone to hallucinations, adversarial attacks or even 'scheming'. The autonomous nature of multi-agentic systems introduces new categories of risk with AI. For one, multi-agent reasoning could introduce a large attack surface for agentic AI hallucinations. It's one thing for an LLM to hallucinate a fact in a blog. How trustworthy would an AI agent be if it hallucinated a day trading strategy that impacts thousands of brokerage accounts? The risks could be even greater. For instance, long-term planning agents (LTPAs) could pose major risks as their ultimate goals may not align with human values. LTPAs are agentic AI systems that are capable of achieving goals over extended timelines. An LTPA's complex, unpredictable strategies may even surpass human oversight. They pose certain basic risks when their objectives could lead to harmful outcomes – they may optimize their objectives in ways that ignore ethical boundaries such as due regard for the environment. LTPAs could develop detrimental secondary objectives such as prioritizing their own existence and acquiring resources to ensure self-preservation and even engage in resource competition with humans.

It is evident that without proper guardrails, agentic AI systems may generate biased outputs, violate ethical guidelines or create severe unintended consequences. That's where AI guardrails come in. These are the technical tools, processes and governance mechanisms companies deploy to ensure their systems conform to evolving policies on AI and responsible practices with AI.

The architecture for implementing AI guardrails may typically start with user query processing, comprehensive safety frameworks with denied topics filters, and sophisticated content screening mechanisms that check for inappropriate material or harmful instructions.



Case in point:

AI chip pioneer NVIDIA recently introduced new inference microservices for AI guardrails, addressing key enterprise concerns. These include content safety controls trained on over 35,000 human-annotated data samples, topic control mechanisms to keep agents focused on approved subjects, and jailbreak detection to prevent system manipulation. The company notes that, "..one in ten organizations are already using AI agents today, and more than 80% plan to adopt AI agents within the next three years".

However, successful agentic AI implementation requires more than technological capability – it demands a comprehensive approach to risk management, ethical deployment, and human-AI collaboration. Here are the high-level areas that enterprises must base their AI guardrails initiatives on.

Multi-Layered Technical Controls

Put technology-based controls in place to handle exceptions at multiple levels. Infra giant Amazon Web Services recommends a multi-layered approach to AI guardrails including prompt templates that provide blueprint structures for input and output, tone and domain specifications through system prompts, and external validation checks and filters. Advanced implementations could include dynamic guardrail systems that self-evolve security measures, multimodal safety mechanisms across text, images, audio, and video, and sophisticated privacy protection with real-time anonymization.

Governance and Risk Assessment

Comprehensive AI governance goes beyond technology features. Organizations should establish AI councils requiring safety reviews for all new AI projects, ensuring the business processes the projects encompass protect against AI risks. Begin by assessing which tasks and workflows are well-suited for agentic AI, map potential risks, and create mitigation plans starting with low-risk use cases, using non-critical data.

Operational Resilience

An operations-focused approach to responsible agentic AI may choose priorities such as accuracy through constraints like topic classification, safety through toxicity detection mechanisms, transparency with clear disclosure patterns, or empowerment through meaningful human-AI hand-offs.

How do enterprises start the journey towards being stewards of responsible AI deployments?

Embrace Measured Adoption

Experts warn that most organizations aren't agent-ready. Successful implementation requires exposing enterprise APIs and ensuring enterprise readiness beyond model capabilities. Organizations should focus on infrastructure preparation, data quality, and system integration capabilities.

Invest in Observability

Industry veterans call for buildouts of trustworthy agentic AI through observability tools, behavioral guardrails, and robust evaluation methods including tracing decision paths and logging internal states.

Plan for Compliance

Strong compliance frameworks are vital for scaling agentic AI systems while maintaining accountability. Organizations must balance innovation speed with responsibility, particularly in regulated industries.

As 2025 rolls on, organizations that successfully balance the transformative potential of agentic AI with robust guardrails ensuring safety, ethics, and resilience will see more successful and secure deployments. Businesses that master this balance will gain a competitive edge in the marketplace.

For more information, contact infosysbpm@infosys.com

Infosys®
Navigate your next