McKinsey
& Company

McKinsey Explainers

What are AI guardrails?

AI guardrails help ensure that an organization's AI tools, and their application in the business, reflect the organization's standards, policies, and values.

November 2024

You know about guardrails on the highway: barriers along the edge of the road that protect vehicles from veering off course and into danger. With the advent of generative AI (gen AI), the concept of guardrails also applies to systems designed to ensure that a company's AI tools, especially large language models (LLMs), work in alignment with organizational standards, policies, and values.

While gen AI can improve a company's efficiency, innovation, and competitive advantage, it can also introduce challenges and risks. As companies' adoption of gen AI increases rapidly, guardrails are crucial to the responsible use of AI. Guardrails can identify and remove inaccurate content that's generated by LLMs, as well as monitor and filter out risky prompts. Such risky content can include security vulnerabilities, hallucinations, toxic or inappropriate content, and misinformation.

But just as guardrails on the highway don't eliminate the risk of injuries or fatalities, AI guardrails don't guarantee that AI systems will be completely safe, fair, compliant, and ethical. For the best results, companies can implement AI guardrails along with other procedural controls (for example, AI trust frameworks, monitoring and compliance software, testing and evaluation practices), as well as a proper AI operations technology stack, which scales the governance of AI across an organization.

What are the benefits of AI guardrails?

To create the right environment for gen AI innovation and transformation, organizations should ensure that the technology can operate safely and responsibly—with AI guardrails playing a critical role. Here are a few benefits that guardrails can offer an organization as it implements AI:

— *Privacy and security.* AI systems are susceptible to attacks from malicious actors who exploit vulnerabilities to manipulate AI-generated outcomes. Guardrails can shore up AI systems against such attacks, helping to protect an organization and its customers.

— *Regulatory compliance.* With increasing government scrutiny of AI, organizations need to ensure that their AI systems comply with existing and emerging laws and standards. By helping a company maintain its gen AI compliance, guardrails can mitigate the risk of legal penalties and liabilities from the use of these tools.

— *Trust.* Maintaining trust with customers and the broader public is paramount for organizations. Guardrails enable continuous monitoring and review of AI-generated outputs, which can reduce the risk of errant content being released outside of the company.

What are the main types of AI guardrails?

Guardrails are grouped according to their purpose and the types of risks they address. (For more information about our methodology for creating guardrails, see sidebar, "What is HyPe?") McKinsey has developed a taxonomy of guardrails, based on specific risks:

— *Appropriateness* guardrails check if the content generated by AI is toxic, harmful, biased, or based on stereotypes and filter out any such inappropriate content before it reaches customers.

— *Hallucination* guardrails ensure that AI-generated content doesn't contain information that is factually wrong or misleading.

— *Regulatory-compliance* guardrails validate that generated content meets regulatory requirements, whether those requirements are general or specific to the industry or use case.

— *Alignment* guardrails ensure that generated content aligns with user expectations and doesn't drift away from its main purpose. These guardrails can help maintain brand consistency, for example.

What is HyPe?

HyPe is a set of technology-agnostic, adaptable components developed by McKinsey’s CustomerOne service line and powered by QuantumBlack, AI by McKinsey. Organizations can deploy HyPe components to accelerate hyperpersonalized customer engagements such as marketing messages. As part of the HyPe development process, we held workshops with multiple stakeholders to discuss the risks associated with hyperpersonalized content developed by AI as well as strategies to mitigate these risks. These workshops, along with analysis of the performance of the large language model used during the building and testing phases of HyPe, allowed us to create the guardrail categories described in this article.

- *Validation* guardrails check that generated content meets specific criteria: that is, the content contains or does not contain certain information. If a piece of generated content is flagged by a validation guardrail, the content can be funneled into a correction loop to fix the error. Validation should be the last of a series of tasks that guardrails perform. After that, a human validator should review flagged or ambiguous cases that require human reasoning.

A variety of open-source libraries have been developed so organizations can add guardrails to their AI systems easily. The machine learning platform Hugging Face has released the Chatbot Guardrails Arena, which stress-tests LLMs and privacy guardrails to prevent leaks of sensitive data. Nvidia has built NeMo Guardrails, an open-source tool kit for adding programmable guardrails to LLM-based applications. Guardrails AI is a similar open-source package. LangChain, a framework for developing applications powered by LLMs, also provides a guardrails library to help organizations quickly plug guardrails into the sequence of operations. There are also proprietary tools, such as OpenAI’s Moderation, that analyze text generated by AI models to detect and filter out harmful, inappropriate, or unsafe content, according to predefined categories. Microsoft has developed a similar guardrail to monitor chatbot-generated content for Azure, its suite of AI services.

How do guardrails work?

Guardrails are built using a variety of techniques, from rule-based systems to LLMs. Ultimately, though, most guardrails are fully deterministic, meaning the systems always produce the same output for the same input, with no randomness or variability. Generally, guardrails monitor AI systems’ output by performing a range of tasks: for example, classification, semantic validation, detection of personally identifiable information leaks, and identification of harmful content. To perform these tasks, AI guardrails are made up of four interrelated components, each of which plays a crucial role:

- *Checker.* The checker scans AI-generated content to detect errors and flag issues, such as offensive language or biased responses. It acts as the first line of defense, identifying potential problems before they can cause harm or violate ethical guidelines.

- *Corrector.* Once the checker identifies an issue, the corrector refines, corrects, and/or improves the AI’s output as needed. It can correct inaccuracies, remove inappropriate content, and ensure that the response is both precise and aligned with the intended message. The corrector works iteratively, refining the content until it meets the required standards.

- *Rail.* The rail manages the interaction between the checker and corrector. It runs checks on the content and, if the content fails to meet any standard, triggers the corrector to make adjustments. This process is repeated until the content passes all checks or reaches a predefined correction limit. The rail also logs the processes of the checker and corrector, providing data for further analysis.

— *Guard.* The guard interacts with all three of the other components, initiating checkers and correctors along with rails, coordinating and managing rails, aggregating the results from rails, and delivering corrected messages.

When designing guardrails, organizations should ensure that they can be easily integrated with their existing technology stacks and customizable to meet the needs of different use cases.

AI agents are also emerging as tools that can function as guardrails. Organizations can use AI agents to automatically check and correct LLM-produced content that has been flagged by guardrails. Early models of AI agents can autonomously monitor, adjust, and regulate AI-generated outputs, as other AI guardrails can do.

How can AI guardrails generate value?

AI guardrails are not only a tool for meeting compliance or ethical requirements; they can also help create a competitive advantage. For one thing, guardrails can help organizations build trust with their customers and avoid costly legal issues. They can also help organizations use AI more responsibly, and thereby attract and retain top talent.

To maximize the potential for value creation, organizations can scale their AI guardrails by embedding them into enterprise platforms. Iguazio by McKinsey provides AI guardrails in the production environment to help ensure AI governance at scale and reduce the risks of data privacy breaches, bias, hallucinations, and intellectual property infringement.

The financial-services company ING offers an example of how an organization can create value with AI guardrails. ING developed an AI chatbot with guardrails to ensure accurate and safe customer interactions. The guardrails were applied to filter out sensitive information and potentially risky advice to customers, as well as ensure compliance. Since this was a customer support tool, it was vital from the beginning to design the AI chatbot with guardrails to ensure it provided safe outputs to ING's customers while also complying with regulatory standards.

How can an organization deploy AI guardrails at scale?

Here are some initial, high-level steps that companies can take:

— *Design the guardrails with multidisciplinary teams.* Work with diverse stakeholders, including legal teams, to build guardrails based on the actual risks and effects that might stem from AI.

— *Define content quality metrics.* These metrics should be tailored to the desired content outputs and based on specific business goals, standards, and regulations. They could include factors such as offensiveness, bias, and alignment with brand guidelines.

— *Take a modular approach.* Build guardrail components that are reconfigurable for different gen AI use cases and can be easily embedded—and also scaled—in the company's existing systems.

— *Adopt a dynamic approach.* Gen AI tools are probabilistic systems that dynamically adjust their outputs based on user-generated inputs. This means that the same input might not always result in exactly the same output, which can sometimes be a problem. An organization should put in place rule-based guardrails with dynamic baselines for a model's outputs that can change based on different variables.

— *Steer with existing regulatory frameworks.* Use existing and emerging regulatory, legal, and compliance frameworks, as well as industry best practices, to create "goals" for the guardrails to hit. These can all be used as metrics against which companies can measure their models' performance.

- *Develop new capabilities and roles.* Upskill a new generation of practitioners who are accountable for the models' outcomes and for ensuring AI transparency, governance, and fairness—by, for example, embedding documentation, accountability, and compliance processes into organizations' ways of working with AI-based tools.

The rapid rise of AI has complicated the compliance picture for companies of all stripes working in, and with, technology. Guardrails can help companies get ahead of related risks and create a safer space for gen-AI-related innovation and transformation. For example, organizations could apply AI guardrails to product development, where active safety testing is a critical step. Product development processes, which are typically owned by product leaders or engineers, would need to become more multidisciplinary to incorporate the perspectives of ethicists, as well as leaders in compliance, risk, and operations. While it might seem that all of these steps and changes could slow things down for a company, they are actually designed to help organizations better manage AI-related crises and, hopefully, avert them altogether. Moving forward, we can expect not only new types of AI systems but also new standards for how these systems are developed and operationalized.

Articles referenced:

- "The state of AI in early 2024: Gen AI adoption spikes and starts to generate value," May 30, 2024, Alex Singla, Alexander Sukharevsky, Lareina Yee, and Michael Chui, with Bryce Hall
- "Implementing generative AI with speed and safety," *McKinsey Quarterly*, March 13, 2024, Oliver Bevan, Michael Chui, Ida Kristensen, Brittany Presten, and Lareina Yee

Get to know and directly engage with senior McKinsey experts on AI guardrails

Lareina Yee is a senior partner in McKinsey's Bay Area office, where **Roger Roberts** is a partner; **Mara Pometti** is a consultant in the London office; and **Stephen Xu** is a senior director of product management in the Toronto office.

Designed by McKinsey Global Publishing