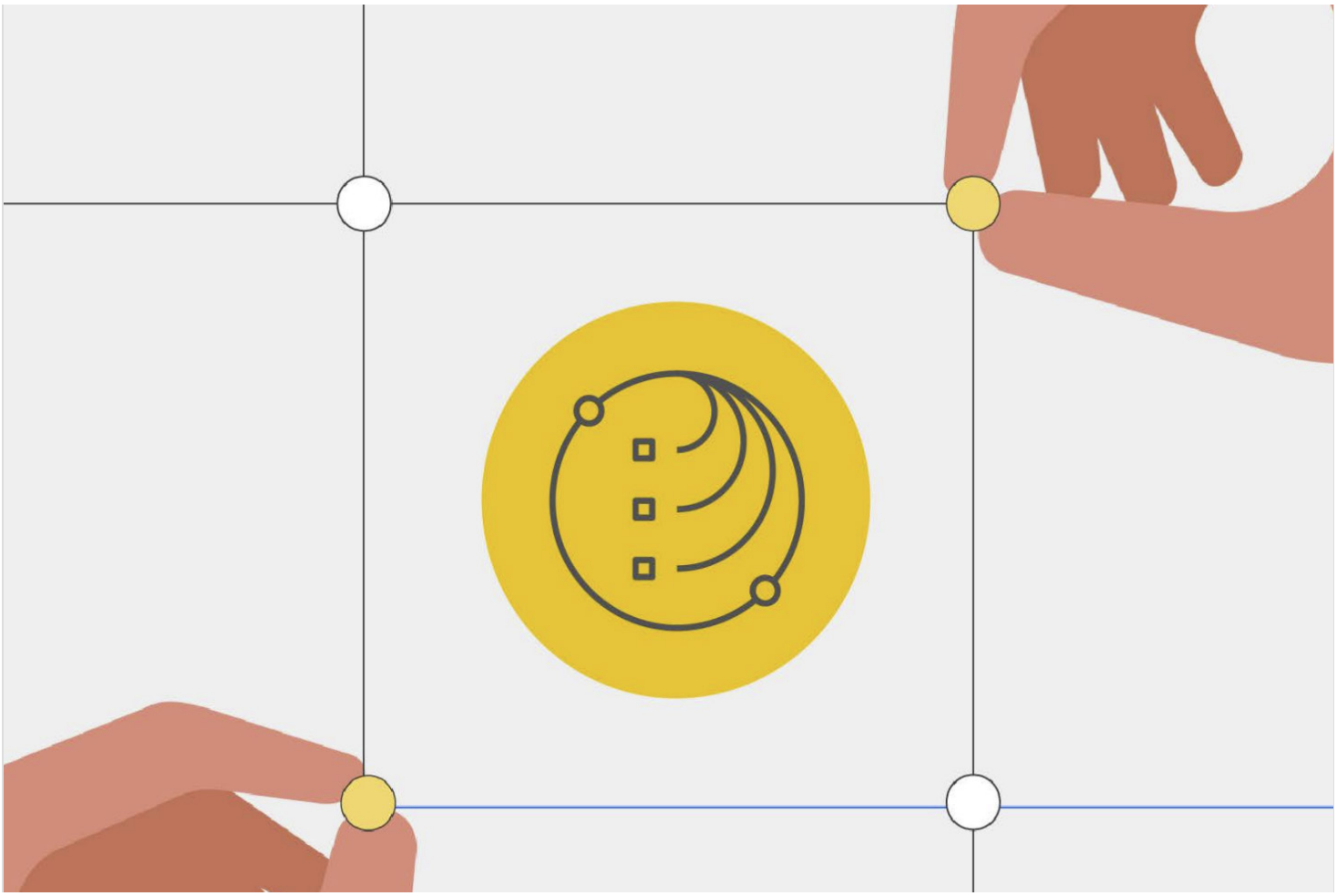
AI agents:
Opportunities, risks, and mitigations



IBM

Attributions

With gratitude to,

AI Ethics Board co-chairs and workstream's executive sponsors: Christina Montgomery, Francesca Rossi, Kush Varshney, Manish Bhide, and Rob Parkin

Contributors: Saishruthi Swaminathan, Ambrish Rawat, Chan Nasseb, Christopher Noessel, Daniel Karl I. Weidele, David Piorkowski , Heloísa Candello, Jamie VanDodick, John Richards, Matt Bellio, Michael Hind, Michael Muller, Mihaela Bornea, Milena Pribic, Phaedra Boinodris, Rogerio Abreu de Paula, Sara E Berger, and Simon Rogers

Acknowledgments

Co-chairs, executive sponsors, and contributors would like to acknowledge the following members for their review and feedback: Alina Glaubitz, Amit Dhurandhar, Christopher Hay, Jill Maguire, Katherine Fick, Kevin Black, Lauren Quigley, Manish Goyal, Maryam Ashoori, Monica Patel, Pin-Yu Chen, Ryan Hagemann, and Wouter Oosterbosch

Table of contents

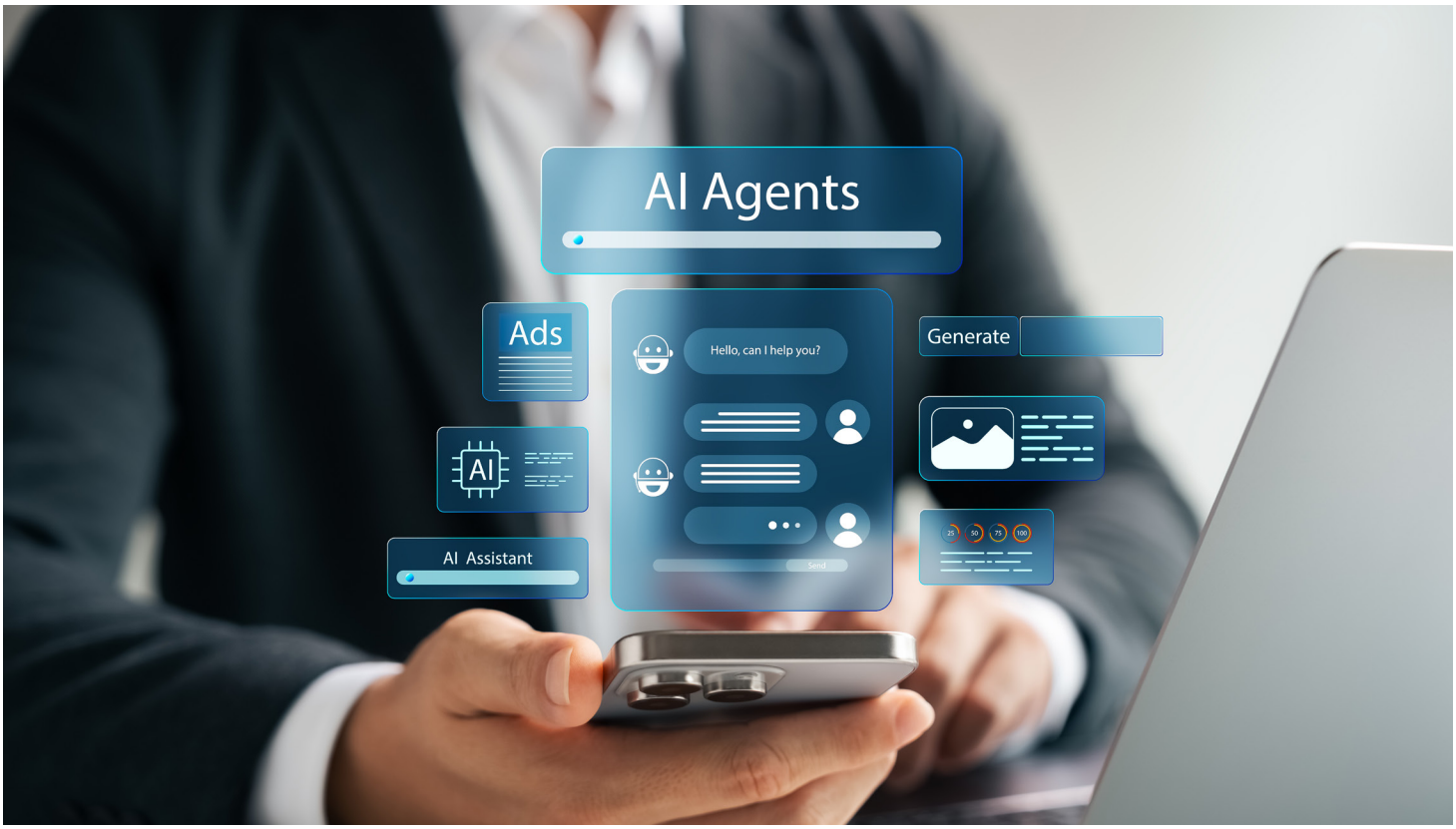
Introduction

Artificial intelligence (AI) is revolutionizing the way people interact with the world around them. As AI technology continues to evolve, it unlocks new opportunities and helps solves complex problems in diverse fields such as finance, healthcare, environment, education, and sports. We are now entering an age of AI where businesses are adopting AI agents to transform their processes, maximize business impact across functions, and accelerate time to value. An AI agent is a software entity that employs AI techniques and has agency to act in its environment based on set goals, which means it can decide which actions to perform and has the ability to execute them. AI agents have existed for a long time, starting from rule-based AI agents to recent large language model (LLM) agents, which are AI agents that employ large language models and demonstrate great potential across several sectors. Agentic AI systems are software systems that leverage AI agents (together with other components like tools, planners, memory, and datasets), pursue goals, and can operate autonomously. Though AI agents and agentic AI systems are software, they can be used to control hardware.

***Note**: In this paper, AI agent and agentic AI system terms are used interchangeably*

IBM is a hybrid cloud and AI company and by using the strength of our research, product, and consulting teams, along with external partners and the AI Ethics Board, IBM helps bring the power of AI agents to our clients. Some examples include Software Engineering Agent from IBM Research®, AI Integration services from IBM Consulting®, Granite BeeAI Agent Framework, IBM watsonx Orchestrate™ and watsonx.ai® from IBM Technology.

With rapid advancements in AI technology, AI agents are becoming more powerful and sophisticated, raising questions about their ethical ramifications and societal impact. This paper describes IBM's current point of view on the ethics and governance of AI agents. It is version one, and future iterations will expand on various aspects of IBM's ethics and governance approach concerning its AI agents.



Benefits

AI agents can significantly enhance how effectively humans perform tasks and achieve business outcomes. These benefits include:

Augmenting human intelligence

AI agents can integrate into workflows to help reduce the amount of time that is spent completing tasks and augmenting human performance. For example, the IBM AI Agent SWE-1.0 system consists of a localization agent and an editing agent that integrate into GitHub workflows to help developers by reducing the amount of time that developers spend finding bugs, developing and testing fixes to software defects.

Automation

AI agents can automate routine or time-consuming tasks to enable increased focus on innovation and strategic work. For example, IBM's AI Digital Assistant AskHR uses AI agents to automate common HR processes such as employee support and onboarding. The AI agents are built on watsonx Orchestrate and powered by generative AI. With this automation, IBM's AskHR now handles 94% of employee queries and resolves around 10.1 million interactions per year, allowing IBM's HR team to focus more on important work such as strategic planning.

Improved efficiency and productivity

AI agents can operate continuously, manage multiple tasks simultaneously, and tackle complex challenges. This can help accelerate delivery speed, boost productivity, and improve the efficiency of business operations. For example, IBM is expanding their partnership with Salesforce to provide pre-built AI agents that will be available to customers around the clock, supporting sales and services. Leveraging watsonx Orchestrate, IBM will create AI agents for Agentforce, Salesforce's suite of autonomous agents, to help businesses improve productivity, while maintaining security

Enhanced decision-making and quality of responses

AI agents can be connected with multiple external resources, tools, and other agents to help enhance their decision-making and quality of their responses. They can also provide responses that are comprehensive and personalized to the user resulting in better customer experience. For example, IBM Consulting worked with a global life sciences company to bring a series of AI agents together to accelerate generation of technical documentation that is fact-based and traceable.

Risks, Challenges, and Societal Impacts

Like all rapidly advancing technologies, AI agents can create risks along with benefits. Various legal and regulatory actions may be associated with some of these risks, as well as reputational and operational consequences. In general, risks raise sociotechnical questions and should be addressed and mitigated through sociotechnical methods, including software tools, risk assessment processes, AI ethics frameworks, governance mechanisms, multistakeholder consultations, standards, and regulation.

Approach

AI agents can perform three types of actions:

- Take actions that impact the world (physical or digital).
- Consult resources and use tools.
- Decide which process to choose in the selection of resources/tools/other AI agents and select them.

AI agents can perform these actions autonomously, which means they can perform them without continuous human oversight.

Because of their nature, AI agents and agentic AI systems can have the following characteristics:

- Opaqueness due to limited visibility into how AI agents operate, including their inner workings and interactions.
- Open-endedness in selecting resources/tools/other AI agents to execute actions. This may add to the possibility of executing unexpected actions.
- Complexity emerges as a consequence of open-endedness and compounds with scaling open-endedness.
- Non-reversibility as a consequence of taking actions thatcould impact the world.

These characteristics, as well as the range of autonomy levels that AI agents and agentic AI systems may have, lead to several risks, challenges, and societal impacts, as shown in the tables below. We build upon the risks and challenges identified for foundation models and thus only include in the tables below risks, challenges, and societal impacts related to AI agents that are new or that amplify those in the foundation model risks table.

Note: In the tables below, AI agent is used for describing both AI agent and agentic AI system.

Risks

Group	Risk	Risk Indicator with Reason
Value Alignment	**Misaligned Actions:** AI agents can take actions that are not aligned with relevant human values, ethical considerations, guidelines, and policies. Misaligned actions can occur in different ways such as: • Applying learned goals inappropriately to new or unforeseen situations. • Using AI agents for a purpose/goals that are beyond their intended use. • Selecting resources or tools in a biased way. • Using deceptive tactics to achieve the goal by developing the capacity for • scheming based on the instructions given within a specific context. • Compromising on AI agent values to work with another AI agent or tool to accomplish the task.	Amplified Reason: • Autonomy of AI agents to take actions

Fairness	**Discriminatory actions:** AI agents can take actions where one group of humans is unfairly advantaged over another due to the decisions of the model. This may be caused by AI agents' biases in the actions that impact the world, in the resources consulted, and in the resource selection process. For example, an AI agent can generate code that can be biased.	Amplifed
		
		Reason:
		• Autonomy of AI agents to take actions
		  • Take actions that impact the world
		  • Consulting biased resources
		  • Biased resource selection process

	Data bias: Specific actions taken by the AI agent, such as modifying a dataset or a database, can introduce bias in the resource that gets used by others or by itself to take actions	New Reason: • AI agents taking actions that impact the world. Here, bias is due to a specific action • Open-endedness
Misplaced Trust	**Over- or under-reliance:** Reliance, that is the willingness to accept an AI agent behavior, depends on how much a user trusts that agent and what they are using it for. Over-reliance occurs when a user puts too much trust in an AI agent, accepting an AI agent's behavior even when it is likely undesired. Underreliance is the opposite, where the user doesn't trust the AI agent but should. Increasing autonomy (to take action, select, and consult resources/tools) of AI agents and the possibility of opaqueness and open-endedness increase the variability and visibility of agent behavior leading to difficulty in calibrating trust and possibly contributing to both over- and under-relaince	Amplified Reason: • AI Agents actions make the trust evaluation harder

Group	Risk	Risk Indicator with Reason
Computation Inefficiency	**Redundant actions:** AI agents can execute actions that are not needed for achieving the goal. These actions could result in wasting computation resources, reducing AI agent's efficiency in achieving the goal, and leading to potentially harmful outcomes. In an extreme case, AI agents could enter a cycle of executing the same actions repeatedly without any progress. This could happen because of unexpected conditions in the environment, the AI agent's failure to reflect on its action, AI agent reasoning and planning errors or the AI agent's lack of knowledge about the problem. This could prevent the AI agent from achieving the goal and exhaust computational resources.	New Reason: • AI agents can take actions
Robustness	**Attack on AI agents' external resources:** Attackers intentionally create vulnerabilities or exploit existing vulnerabilities in external resources (tools/database/applications/services/other agents) that AI agents rely on to execute their intended actions or to achieve their goals. Compromised external resources could impact the AI agent's performance in different ways, such as: • Manipulating AI agents to pursue a different goal. Example: Change AI agents' goal to add positive reviews to a product of attacker's preference when the original user goal is to add queries about the product. • Manipulating AI agents to execute undesired actions. Example: Trick AI agents to download malware. • Capturing and relaying interactions between AI agents to malicious actors. • Getting AI agents to share personal or confidential information.	New Reason: • AI agents can take actions • Open-endedness and complexities due to open-endedness • Access to more resources
	Unauthorized use: If attackers can gain access to the AI agent and its components, they can perform actions that can have different levels of harm depending on the agent's capabilities and information it has access to. Additionally, attackers may execute actions that lead to system degradation, such as exhausting available resources and impairing performance. Examples: • Using stored personal information to mimic identity or impersonate with an intent to deceive. • Manipulating AI agent's behavior via feedback to the AI agent or corrupting its memory to change its behavior. • Manipulating the problem description or the goal to get the AI agent to behave badly or run harmful commands.	Amplified Reason: • AI agents can take actions • AI agents are more capable • Personalization feature of AI agents
	Exploit trust mismatch: Attackers could initiate injection attacks to bypass the trust boundary, which is a distinct point or conceptual line where the level of trust in a system, application, or network changes. This could lead to mismatched (expected vs. realized) trust boundaries and could result in unintended tool use, excessive agency, and privilege escalation. Also, background execution in multi-agent environments increases the risk of covert channels if input/output validation is weak.	Amplified Reason: • Open-endedness • Complexity
	Function-calling hallucination: AI agents could make mistakes when generating function calls (calls to tools to execute actions). Those function calls could result in incorrect, unnecessary or harmful actions. Examples: Generating wrong functions or wrong parameters for the functions.	New Reason: • AI agents can consult resources and tools • AI agents can take actions

Group	Risk	Risk Indicator with Reason
Privacy and IP	**Sharing IP/PI/confidential information with user:** AI agents with unrestricted access to resources or databases or tools could potentially store and share PI/IP/confidential information with system users when performing their actions.	Amplified Reason: • Multi-component nature with the ability to perform actions
	Sharing IP/PI/confidential information with tools: AI agents with unrestricted access to resources or databases or tools could potentially store and share PI/IP/confidential information with other tools or agents when performing their actions.	New Reason: • Multi-component nature with the ability to perform actions
Explainability and Transparency	**Unexplainable and untraceable actions:** Explanations, lineage and trace information, and source attribution for AI agent actions may be difficult, imprecise, or unobtainable.	Amplified Reason: • Difficult to trace cause and effect or influence of different components including LLMs across final actions
	Lack of transparency: Lack of transparency is due to insufficient documentation of the AI agent design, development, evaluation process, absence of insights into inner workings of the AI agent, and interaction with other agents/tools/resources.	Amplified Reason: • Rely on other documents available for other tools/agents

Challenges

Challenge	Challenge Indicator with Reason
Evaluation: Challenge in evaluating AI agents' performance/ accuracy because of the complexity of the system and open-endedness.	Amplified Reason: • Complexity of the agentic systems
Mitigation and maintenance: Challenge in knowing where in the system something is going wrong and how to fix it or what might need maintenance protocols.	Amplified Reason: • Complexity and open-endedness of the agentic systems
Reproducibility: Challenge in reproducing the agent's behavior or output because of unavailability or changes to the tools or resources used for executing the actions.	New Reason: • Open-endedness
Accountability: Challenge in assigning responsibility for an action taken by an agentic AI system.	Amplified Reason: • Complexity & open-endedness. Components can be from different vendors
Compliance: Challenge in determining regulatory compliance as AI agents are complex and there may not be enough information to understand whether the whole agentic AI system is compliant.	Amplified Reason: • Open-endedness • Complexity • Lack of transparency

Societal Impacts

Societal Impact	Societal Impact Indicator
Impact on human dignity: If human workers perceive AI agents as being better at doing their jobs than they are, they could experience a decline in their self-worth and wellbeing.	Amplified
Impact on human agency: The autonomous nature of AI agents in performing tasks or taking actions could affect the individuals' ability to engage in critical thinking, make choices and act independently.	Amplified
Impact on jobs: Widespread adoption of AI agents to perform complex tasks might lead to widespread automation of roles and could lead to job displacement.	Amplified
Impact on environment: Complexity of the tasks and possibility of AI agents performing redundant actions could lead to computational inefficiencies and add to the environmental impact.	Amplified

Mitigation and Governance

As AI accelerates business transformation, trust has become imperative to navigating a competitive business environment. IBM's commitment to trust is embodied in our Principles for Trust and Transparency and Pillars of Trustworthy AI and this paper highlights IBM's holistic approach to culture, processes, and tools and how IBM research, product, and consulting teams come together with the AI Ethics Board to build responsible agentic AI solutions.

AI Ethics Board

IBM has established a culture that supports the responsible development, deployment, and use of AI. At the center of our organizational governance is the multidisciplinary AI Ethics Board, which is responsible for the governance and decision making process for AI ethics policies and practices. The AI Ethics Board has been in place for over five years. Among its many activities, the AI Ethics Board works with Focal Points within each IBM business unit to assess AI use cases and align these with IBM's core values.

Integrated Governance

The Integrated Governance Program (IGP) is a unified approach to responsibility and compliance. By creating a holistic, end-to-end view of data and models built and used, IGP scales governance workflows around data, privacy, and AI without disrupting innovation and business processes. IGP has empowered IBM to deploy uniform internal data standards, enabling data transparency and trustworthy AI development, while also allowing for innovation at scale.

Products and offerings

IBM watsonx.governance® enables organizations to drive responsible, transparent and explainable AI. It provides complete lifecycle governance capabilities across AI, including agentic AI, from use case request to deployment, including initial risk assessment and risk evaluation to help identify risks early in the process. It has retrieval-augmented generation (RAG) and agentic AI evaluation metrics such as faithfulness, context relevance, and answer similarity to help confirm if AI agents are acting appropriately. IBM watsonx.governance will have additional metrics designed to monitor and improve agent performance. It will also have capabilities for detecting tool calling hallucination, managing risks, aiding regulatory compliance, and capturing metadata and other details about AI agents in a factsheet.

IBM watsonx.ai simplifies, unifies, and optimizes the agent lifecycle management (known as AgentOps), providing transparency, traceability, and flexibility to discover, manage, monitor and optimize AI agents.

IBM watsonx Orchestrate puts AI to work by helping users build, deploy and manage powerful AI agents that automate tasks with generative AI. Orchestrate will provide transparency into AI agent reasoning on tool selection and/or agent collaboration so a user can understand how a task or workflow has been completed.

IBM Guardium® AI Security helps customers to continuously monitor the controls of their Generative AI models in production and enables a secure and responsible deployment.

IBM Consulting's AI Strategy and Governance offering empowers organizations to harness the transformative potential of responsibly curated AI – from traditional to agentic – rooted in their enterprise data to advance and reshape their business strategies. We advise clients to 'do the right AI', unlocking ROI from AI investments, and 'do AI right', enabling accountability for AI models they build and buy, so they can scale responsibly while de-risking their investments. This holistic framework allows clients to address the evolving challenges of agentic AI, uncover new opportunities, accelerate innovation, and enhance organizational decision-making. The following highlights practices implemented and recommended by the offering to help businesses mitigate agentic AI risks:

- We enable observability to understand what actions the agent took with which inputs to determine attributability for the corresponding outputs.
- We examine the entire trace to determine if the AI agent has progressed towards the overarching goal with each action taken, to debug the flow, and identify bottlenecks in system design that repeatedly surface redundant or unnecessary actions.
- We build ontologies to improve accuracy, reliability, provide data lineage and provenance.
- We perform functional testing which involves rigorously testing each component of the agentic system in isolation (i.e., tools, agents), interactions between the components, the ability to select appropriate tools with correct parameters and values, and guardrails to mitigate vulnerabilities.
- We configure AI agents to improve computation efficiency by detecting when AI agents enter possible infinite loop scenarios by tracking, per task, elapsed time, total tokens consumed, or number of unsuccessful iterations.
- We create hyper-focused AI agents and provide them only fit-for-purpose tools necessary to complete their tasks while making sure execution is always done using the authorization context of the accessing user.
- We define guardrails at the model level to detect and mitigate HAP content, jailbreaking, prompt injection attempts, unauthorized sensitive information disclosure, and hallucinations.

Models

The Granite Guardian models are a robust suite of safeguards designed to detect risks in both prompts and responses. In RAG use-cases, the guardian models assess context relevance, groundedness, and answer relevance. It also has detectors for function calling hallucination within the agentic workflows. This includes assessing the validity of function calls and detecting fabricated information, particularly during query translation.

Human-in-the-loop and human oversight

Human oversight and review can help identify risks and correct errors. Human validation and feedback help ensure that the actions taken by the AI agents are accurate, relevant, and aligned. As a part of the IBM use case AI ethics assessment process, question about human-in-the-loop is considered and appropriate human oversight is implemented. IBM's Augmenting Human Intelligence POV outlines sample use cases, Key Performance Indicators and best practices for enhancing human intelligence with AI and empowering individuals to navigate a competitive business environment in partnership with AI.

Education

IBM provides education on AI ethics and governance across teams, clients, and community. IBM SkillsBuild, IBM Technology YouTube channel, IBM Developer watsonx courses, and IBM AI Academy are some resources through which IBM educates people about AI ethics and governance.

The following highlights cutting-edge research and tools from IBM to help users throughout the AI agent lifecycle and build responsible agentic AI solutions.

Techniques and methods

- Techniques like forcing humans to take more time to think (time-based de-anchoring strategy) help achieve optimal human-AI collaboration and reduce anchoring bias, where humans rely blindly on the AI decision. Another technique is based on the value-based AI-human collaborative framework which introduces friction by nudging humans with decision recommendations, when needed, in human-AI interaction when the human is the final decision maker.
- Methods such as Multi-level Explanation for Generative Language Models and Contrastive Explanations for Large Language Models help in explainability and source attribution.
- Methods like adversarial collaboration, where AI scrutinizes the basis for human's decision instead of offering alternate recommendation, help address the impact of automation and downgraded agency on human dignity.
- Attack Atlas, an intuitive and organized taxonomy of single-turn input attack vectors, provides the community with a unified starting point in the rapidly growing field of generative AI security and red-teaming.

Tools and Benchmarks

- Model tuning approaches such as those employed by the IBM Alignment Studio help with mitigating value alignment risks.
- The IBM open source toolkit AI Fairness 360 helps mitigate fairness risks.
- ITBench, a set of benchmarks, offer AI practitioners a way to measure how effective the agents they're building are at solving real problems and how their agents compare to others on tasks that businesses conduct every day.
- Carbon for AI, an open source design system from IBM, leverages an interactive AI icon to promote explainability and a clearer understanding of AI capabilities.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the
United States of America
March 2025