Just in Time

AI Agents and Risk Management

April 2026

iason ESSENTIAL SERVICES FOR FINANCIAL INSTITUTIONS

Executive Summary

- The rise of Large Language Models (LLMs) and the Advanced AI systems which follow, introduce increasingly complex and evolving risks.
- **The following work** aims at providing a **structured overview** of these risks and how they scale with architectural complexity. Across AI systems, risks cluster into four structural dimensions: **Data Risks, Vendor Risks, Architecture Risks, and Human Factor Risks**. While the categories remain consistent, their impact intensifies as systems evolve in complexity (e.g., LLMs to Agents to coordinated Multi-Agent ecosystems), shifting from intrinsic model risks to operational and systemic vulnerabilities.
- New risks, scaling and compounding with AI complexity and deployment, require an **Agent Risk Management (ARM) framework**, designed to extend traditional Model Risk Management (MRM) to agentic and multi-agent AI systems.

Copyright © 2026 - All rights reserved

At a Glance

Keywords: Artificial Intelligence, AI Risk Management, Risk Taxonomy, AI Agents, Large Language Models

Copyright © 2026 - All rights reserved

01

Advanced AI Systems Overview

Next-Generation AI and Autonomous Intelligence Systems

From LLMs to Multi-Agent Systems

Key Milestones in AI Governance

Copyright © 2026 - All rights reserved

Advanced AI Systems Overview 1/3

Next-Generation AI and Autonomous Intelligence Systems

Advances in generative models and reasoning capabilities are now enabling a **new paradigm** in which AI evolve from automation tools to **intelligent collaborators**, capable of supporting complex decision-making and operational processes:

From Rule-Based Automation
Rigid *"if–then"* logic designed for specific tasks. Effective in controlled environments but limited when facing new or dynamic conditions.

To Generative Reasoning
Large Language Models enable **advanced pattern recognition and high-quality content generation**, supporting knowledge work and decision-making.

Toward Autonomous Intelligence
Agent-based and multi-agent systems can plan and execute multi-step objectives, interacting with tools and environments with **limited human intervention**.

New Capabilities Driving Next-Generation AI

Cognitive Reasoning & Planning: AI is evolving from word prediction to structured problem solving, decomposing complex goals into actionable tasks.

Adaptive Learning & Dynamic Adaptability: systems continuously improve and adjust strategies in real time based on interaction and environmental feedback.

Action-Oriented Architectures & Tool Integration: AI is shifting from systems that talk to systems that act, interacting with external tools to execute real-world workflows and operations.

Increasing Need for Risk Management

Advances in computational technologies as high-performance computing and quantum acceleration, will strengthen **model scalability** and enable faster, more accurate decision processes.

As **AI becomes more capable** and is deployed in critical domains, **robust risk management** is essential, addressing unintended behaviors and systemic risks through strong governance.

Key Challenges

The increasing autonomy and sophistication of AI systems also introduce a **set of critical challenges**:

Workforce Impact: managing the effects of AI-driven automation on jobs and skills

Explainability: ensuring outputs and decisions are interpretable and trustworthy

Privacy: safeguarding sensitive data across training, inference, and deployment

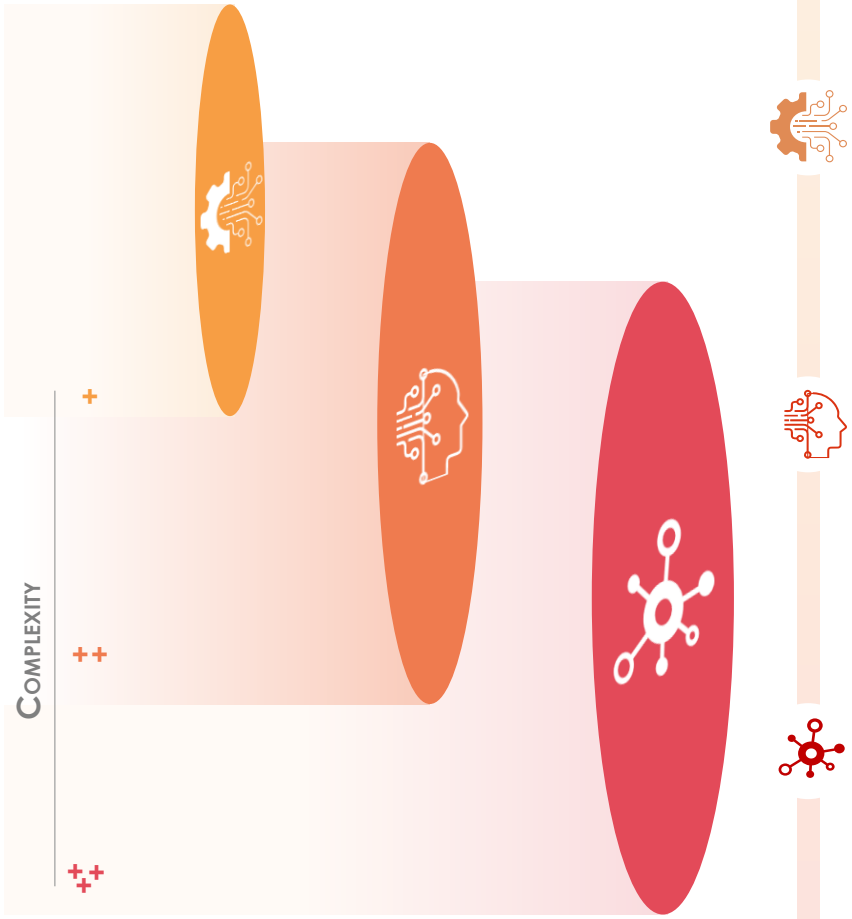
Bias & Fairness: preventing models from inheriting or reinforcing unfair patterns from training data

Safety & Control: guaranteeing systems behave as intended avoiding harmful outcomes

Copyright © 2026 - All rights reserved

Advanced AI Systems Overview 2/3

From LLMs to Multi-Agent Systems



A **Large Language Model*** is an AI system trained on extremely large text datasets to understand, generate, and manipulate natural language. It functions as a single linguistic processing unit capable of answering questions, writing and summarizing text, translating, and generating code.

As a **stand-alone entity** (i.e., ***the foundation model in isolation***) it presents intrinsic limitations, such as:

- it lacks **long-term memory** or any persistent state beyond the inference-time context window;
- it cannot autonomously interact with external tools requiring an external layer for any real-world action
- it has no **intrinsic goals** or no/limited **capacity to plan** complex, multi-step sequences, with apparent reasoning patterns (e.g., Chain-of-Thought) emerging only when elicited through prompting

The LLM evolves from a simple text generator into an **agent** capable of perceiving, reasoning, planning, and acting, through both inputs and external tools.
This transformation is enabled by the integration of key components:

- **Memory**: mechanisms to store and learn from past interactions (short and long-term)
- **Tools**: ability to interact with external systems (e.g., APIs, databases, calculators)
- **Planner/Reasoner**: a module for decomposing goals, sequencing actions, and self-correcting

A **multi-agent system** consists of multiple AI agents, often enhanced LLMs, each with distinct roles, goals, and capabilities interacting with humans, the environment and one another. Within such systems, agents may **collaborate** on shared goals, **coordinate** actions to optimize performance, but they may also **compete** in two specific scenarios:

- **Conflicting objectives**: agents may pursue goals that become mutually incompatible when executed in parallel and advancing one agent's objective reduces system efficiency or directly interferes with another's
- **Limited resources**: agents depend on constrained resources (e.g.; compute, bandwidth, APIs) where one agent's acquisition of a resource limits its availability to others

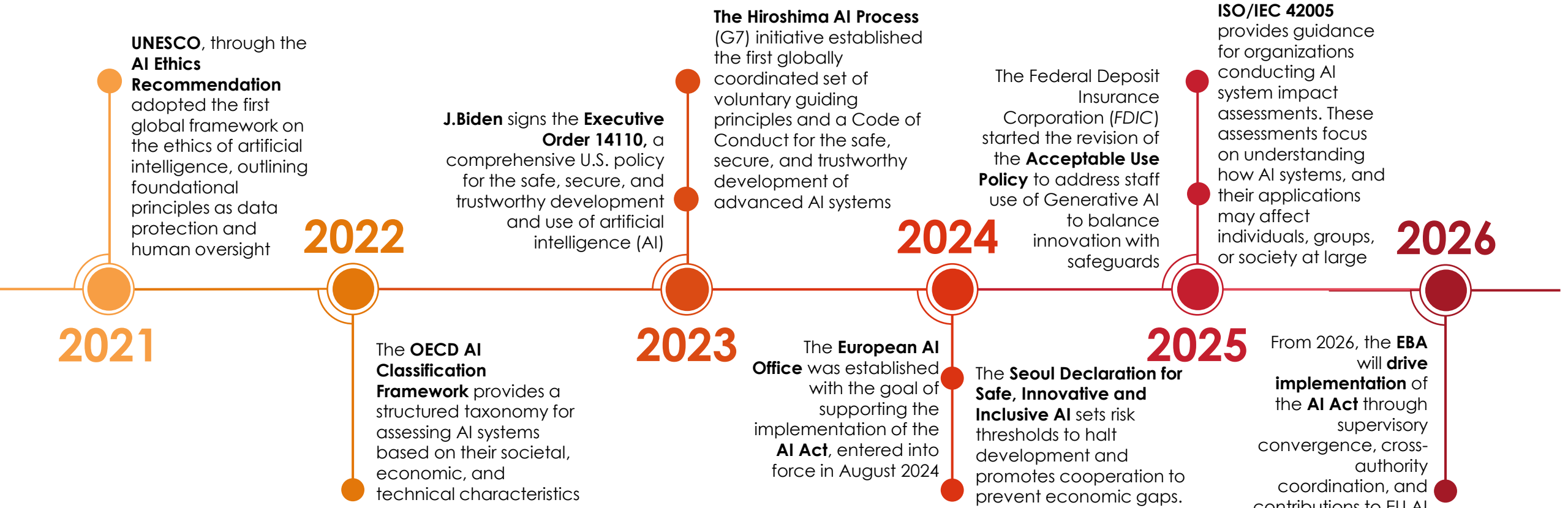
* This definition refers to the stand-alone LLM; the capabilities perceived by users on platforms such as ChatGPT or Gemini derive not from the pure model itself but from the broader agentic framework that surrounds it—providing memory, tool orchestration, task planning, and safety/governance layers.

Copyright © 2026 - All rights reserved

Advanced AI Systems Overview 3/3

Key Global Milestones in AI Governance

As advanced AI systems evolve toward greater autonomy, particularly through agent-based and multi-agent architectures, regulatory frameworks* are still in a **preliminary stage**. While global principles on transparency, accountability, and safety are emerging, the regulatory landscape** is currently in a transitional phase, shifting from broad governance frameworks to more specific oversight mechanisms designed to address the novel risks posed by increasingly autonomous, adaptive, and interconnected AI systems.



* For a deep dive into the AI Regulatory Framework see Artificial Intelligence Act (AI Act) and AI Act: AI System Definition and Prohibited AI Practices
** For a deep dive into the main regulatory milestones see AI: Regulation, Rise and Challenges and USA vs EU: Use and Oversight of AI in Financial Services

Copyright © 2026 - All rights reserved

02

Risks in Advanced AI Systems

Risk Landscape in Advanced AI Systems

LLMs-Specific Risks

Agents-Specific Risks

Risks in Multi-Agents Systems

Risk Amplification Across AI Architectures

Copyright © 2026 - All rights reserved

Risks in Advanced AI Systems 1/5

Risk Landscape Across LLM, Agents and Multi-Agent Systems

LLM-based systems evolve from standalone foundation models to autonomous agents, and ultimately to coordinated multi-agent ecosystems. Across this progression, risks do not fundamentally change in nature — but they increase in **scale, interaction,** and **propagation potential.** Regardless of the specific architecture, risks are clustered into **four structural dimensions**, which apply consistently across all levels of system complexity.

DATA RISKS

Risks related to the quality, reliability, provenance, dynamics, and interpretability of the data that informs or is generated by the system.

VENDOR RISKS

Risks arising from dependencies on external providers, including model updates, API access, pricing volatility, supply chain concentration, and technological lock-in.

ARCHITECTURE RISKS

Risks stemming from system design, component orchestration, configuration complexity, and the interaction mechanisms between models, tools, and agents.

HUMAN FACTOR RISKS

Risks emerging from human interaction with AI systems, including cognitive biases, over-reliance, role confusion, accountability gaps, and organizational misalignment.

How These Risks Scale with System Complexity:

The more complex the AI system, the more previous risks compounds with new ones, amplifying the potential harm they can bring

LLMs

*At the model level, risk is **intrinsic** and **foundational**: it stems from epistemic uncertainty, opaque training data, provider dependency, and probabilistic (non-deterministic) behavior.*

At this stage, risk is embedded in how the model reasons — but it remains relatively contained

Agents

*When LLMs are assembled into an agent, risk turns **operational** and **dynamic**: the system interacts with tools, stores memory, plans and executes actions. Errors can therefore **propagate across steps and over time. Here risk starts scaling.***

Multi-Agents Systems (MAS)

*In multi-agent architectures, risk becomes **systemic** and **spread**: interactions between agents can create unpredictable effects, coordination failures, cascading errors, and blurred accountability.*
*At this level, **risk emerges from the behavior of the ecosystem as a whole.***

Copyright © 2026 - All rights reserved

Risks in Advanced AI Systems 2/5

LLMs-Specific Risks

Large Language Models (LLMs) and **Open Conversational Systems** mark a significant advancement in AI but carry inherent risks which become more pronounced in dynamic, open-ended interactions. **Unlike rule-based chatbots**, which maintain the task-specific operational context, LLMs can retain and use detailed contextual information across turns. This flexibility increases expressive power but reduces predictability and control, heightening **exposure** to **unpredictable** and **domain-specific risks that scale as models are composed into agents and multi-agent architectures**.

Data Risk

Hallucination: LLMs may generate false information from training data or propagate harmful content during inference, contributing to misinformation, poor decisions, reputational damage, or public opinion manipulation.
Focus on generic risks: safety efforts often emphasize disinformation and toxicity while neglecting domain-specific risks, reinforcing the broader "*safety gap*".
Data Contamination: LLMs are often evaluated using generic, domain-agnostic benchmarks that fail to reflect real deployment environments, potentially leading to culturally inappropriate, legally non-compliant, or technically inadequate outputs.

Vendor Risk

Software Supply Chain Vulnerabilities: LLMs rely on complex ecosystems of open-source libraries, model weights, and frameworks; compromising any component can affect the entire system.
Third-Party Dependency Risks: external libraries, APIs, and plugins may introduce hidden vulnerabilities if poorly maintained.
Outdated or Deprecated Components: unpatched libraries, outdated models, or unsupported frameworks exposes LLM systems to known exploits and security weaknesses.

Architectural Risk

Context Window Degradation: in long-context models, extended conversations can reduce precision or cause the model to ignore earlier instructions, undermining reliability and system effectiveness.
Long-term safety challenges: it is inherently difficult to assess safety over extended conversations, where harmful behavior or content may emerge only after complex or lengthy interactions.

Human Risk

Safety hazards: LLMs may generate harmful or unlawful content (e.g.; material related to weapons, adult content, or violations of social norms or specific rules such as toxicity or cultural insensitivity).
Anthropomorphism & Overtrust: the model's fluent tone may lead users to perceive unwarranted authority and follow its advice, even when disclaimers are present and may not meet legal or regulatory standards.

Copyright © 2026 - All rights reserved

Risks in Advanced AI Systems 3/5

Agents-Specific Risks

As AI systems evolve from general LLMs to specialized agents, risks traditionally associated with LLMs, like hallucination or unsafe outputs, manifest differently at the agent level, where autonomy, interaction, and decision-making **amplify** their **impact.**

Data Risk

Quality hazard: concerning the reliability and overall usefulness, reducing effectiveness. They typically manifest as:

- ***Inaccuracy***: responses factually incorrect or misleading;
- ***Lack of robustness***: failures or unstable behavior when faced with unexpected or slightly altered inputs;
- ***Inconsistency***: contradictions within the interaction;
- ***Irrelevance/memory corruption***: outputs that do not address the user's request or, since agents have long-term memory (unlike raw LLMs), a single malicious interaction can "poison" the agent's memory, affecting every future interaction.

Disinformation: production of **inaccurate** or **misleading outputs** due to hallucination arising when the model relies on associative patterns or incomplete context rather than verified information.

Vendor Risk

Model Drift: structural reliance on an **external foundation model** whose updates, safety policies, and behavioral evolution remain outside direct organizational control.

Tool and API Integration Risks: compromised vendor APIs or tools may trigger real-world actions such as executing transactions or modifying data.

Plugin Vulnerabilities: third-party plugins used in agent frameworks can serve as entry points for unauthorized access.

Architectural Risk

Goal Misalignment: as an agent grows in complexity, it may exhibit **unintended capabilities** not anticipated during training pushing the model **beyond** its **intended scope**, making it unpredictable.

Prompt injection and jailbreaking: attacks that manipulate an agent to **bypass its safety constraints**. Malicious inputs can cause the model to generate harmful content, reveal sensitive information, or perform unintended actions.

Human Risk

Safety hazard: the generation of harmful, illegal, or unacceptable content or behavior. manifestations include:

- ***Toxic content*:** outputs containing hate speech, offensive language, threats, or incitement to violence;
- ***Harmful bias and stereotypes*:** reinforcement of discriminatory patterns linked to gender, ethnicity, etc.;
- ***Instructions for dangerous activities*:** guides that enables harmful or illegal activities (e.g.; cyberattacks);
- ***Disclosure of PII*:** revealing personal or sensitive information extracted from training data/ interaction.

Copyright © 2026 - All rights reserved

Risks in Advanced AI Systems 4/5

Risks in Multi-Agents Systems

Risks in multi-agent systems stem from the architectural composition of multiple **autonomous AI agents operating in coordination**. Because these risks originate in the interaction dynamics between agents, they are primarily **structural** and **architectural**.

Data Risk

Information asymmetries & Collusion: when agents hold different information, they may hide data, collude, or act strategically to bypass constraints, undermining coordination and oversight.
Selection pressures: training incentives and deployment rules may favor short-term performance over safety or cooperation, leading agents to develop harmful strategies such as bias amplification, aggression, or destabilizing competition.

Vendor Risk

Systemic Supply Chain Vulnerabilities: if multiple agents rely on the same vendor models, libraries, or frameworks, a single compromised component can propagate across the entire multi-agent ecosystem.
Failure Propagation: vulnerabilities introduced through vendor software can spread through agent-to-agent communication, amplifying system-wide risks.
Heterogeneous Vendor Integration Risks: MAS often integrate components from multiple vendors (models, orchestration frameworks, tools). Interoperability gaps can introduce new security weaknesses.

Architectural Risk

Multi-agent safety: addresses the specific vulnerabilities that arise when many AI agents interact. This includes swarm attacks, heterogeneous attacks, large-scale social engineering, cascading logic failures, and undetectable threats.
Emergent agency: it refers to the rise of higher-level behaviors or capabilities within the collective of agents that are not present individually. This may include the emergence of dangerous goals, such as political manipulation by bot swarms.
Destabilizing dynamics: interactions among adaptive agents can result in complex, unpredictable dynamics, like feedback loops, cyclic behavior, or chaos.
Network effects: interdependencies among agents can cause local errors, biases, or failures to propagate through the system, triggering cascading failures or systemic instability, especially when many agents share the same underlying model.

Human Risk

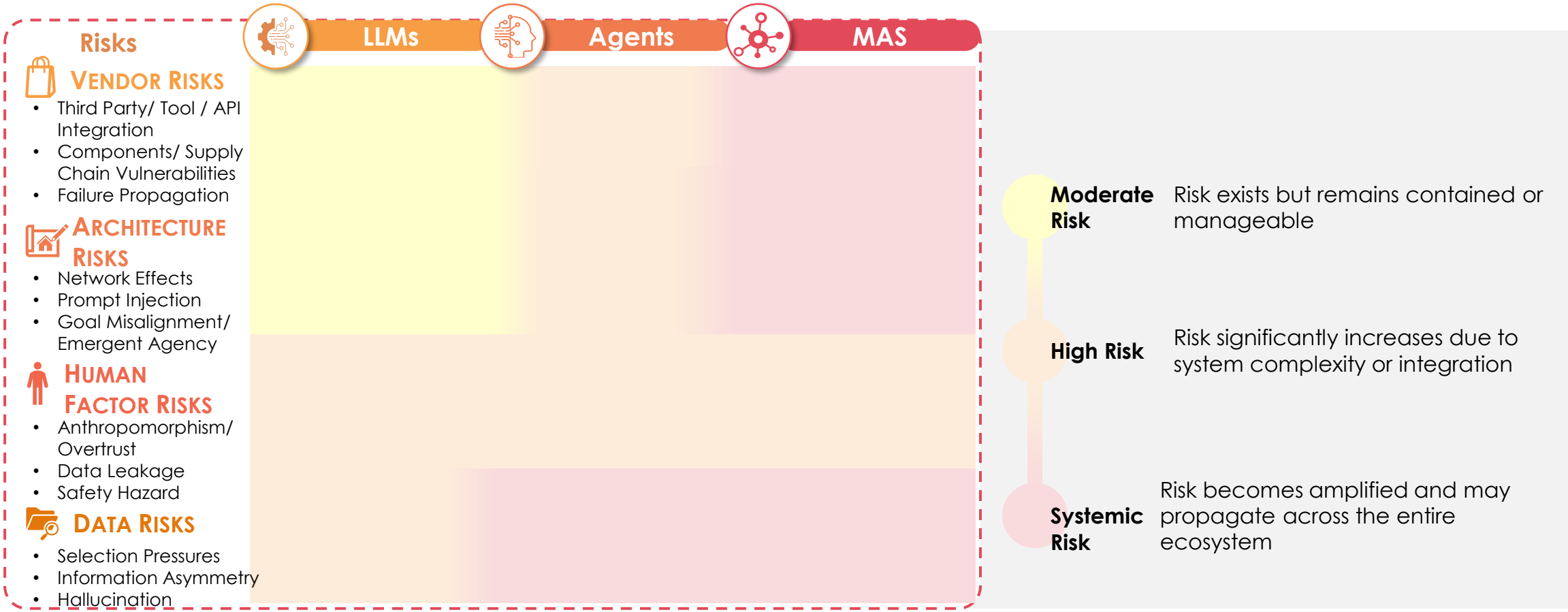
Accountability Opacity: AI agents' ability to make commitments introduces risks when trust mechanisms fail. This can lead to inefficiencies, threats (e.g.; ransomware), or rigid, misaligned commitments.

Copyright © 2026 - All rights reserved

Risks in Advanced AI Systems 5/5

Heatmap of Risk Amplification Across AI Architectures

The heatmap illustrates how risks identified at the LLM level persist and **become amplified as system architectures increase in complexity**, evolving from isolated model risks to systemic multi-agent vulnerabilities.



Copyright © 2026 - All rights reserved

03

Agent Risk Management

Ensuring Control: Why AI Agents Systems Need Governance?

Copyright © 2026 - All rights reserved

Agent Risk Management

Ensuring Control: Why AI Agents Systems Need Governance?

In summary, studying the risks of advanced AI systems is vital to ensure their safe and ethical development and use. The approach must be **proactive**, **holistic**, and **domain-specific**, recognizing that these risks are dynamic and complex, and require constant effort to protect not only users but also organizations and society as a whole.

System complexity and interconnection

Advanced AI, especially in multi-agent setups, operates within **complex networks** where failures can cascade and cause **unpredictable, destabilizing effects**. Risks like feedback loops show that system behavior exceeds the sum of its parts.

Limits of generic safety measures

Standard safeguards, like generic guardrails, often fall short in sensitive domains (e.g., finance). Addressing **domain-specific risks** requires tailored governance and context-aware mitigation strategies.

Need for clear governance and accountability

The inherent complexity of AI systems challenges the clear **attribution of responsibility**. Risk analysis supports strong governance, legal compliance, and continuous improvement, crucial for maintaining trust and avoiding systemic failures.

Emergence of novel risks

AI systems may develop **unexpected behaviors or objectives**, especially when agents interact. Without proactive study, these emergent risks remain undetected and can be exploited or cause harm.

Third-Party Connection Risks

Relying on **external AI providers** can create **vendor lock-in**, tying the organization to their pricing, roadmap, and operational choices. At the same time, effective use of advanced AI requires building **internal skills** for its management, security and maintenance, especially when adopting open-source models.

Social and economic consequences

AI failures can **impact entire organizations or markets**, causing reputational damage, reinforcing social inequalities, or distorting financial systems. These risks go far beyond individual user harm.

Copyright © 2026 - All rights reserved

04

Key Takeaways

Operational and Safety Pillars for AI Agent Governance

Copyright © 2026 - All rights reserved

Key Takeaways

Operational and Safety Pillars for AI Agent Governance

To ensure safe, trustworthy, and context-appropriate deployment of agentic AI systems, it is necessary to formalize an Agent Risk Management (ARM) framework that can integrate seamlessly into existing Model Risk Management (MRM)* and AI governance structures. These elements represent **the key building blocks to extend traditional MRM to govern autonomy-driven and systemic risks in agentic systems**.

Design and implementation

- **Model modification or guardrails:** add filters inside or outside the model to block harmful, biased or unsafe outputs.
- **Network monitoring**: monitor agent interactions to catch errors and preserve stability.
- **Model diversification:** use different model to reduce systemic risk.
- **Secure protocols and adversarial testing**: set rules and simulate attacks.

Contextualization and definition

- **Context-based mitigations:** adjust safety strategies based on use case, users and regulations.
- **Clear risk definitions**: build domain-specific taxonomies to recognize and manage risks effectively.

Monitoring and test

- **Continuous monitoring and updates:** regularly check for harmful behaviors and continuous improvement of the system.
- **Simulation-based testing:** simulations to anticipate system behavior and stress-test before real-world use.
- **Explore emergent behavior:** test in varied settings to identify unexpected outcomes or capabilities in the system.

Governance and process

- **Governance frameworks**: clear processes and responsibilities to manage AI risks and ensure accountability.
- **Context-aware risk approaches**: consider social, legal and domain specific factors in risk assessment.
- **Multidisciplinary set of skilled professionals:** domain experts, technical teams and control functions are required to ensure compliance, coordinated oversight, and safe deployment.
- **Human in the Loop**: keep human control over critical decisions to prevent over-reliance on AI.

*For a deep-dive into GenAI Model Risk Management see GenAI Model Risk Management and Governance in Financial Services - From Principles to Practice and Model Risk Management of GenAI Workflows

Copyright © 2026 - All rights reserved

ESSENTIAL SERVICES FOR FINANCIAL INSTITUTIONS

iason is an international consulting firm that has been supporting both financial institutions and regulators in topics related to Risk Management, Finance and ICT since 2008

Strategy

Strategic advisory on the **design** of **advanced frameworks** and **solutions** to fulfil both **business** and **regulatory needs** in Risk Management and IT departments

Methodology & Governance

Implementation of the designed **solutions** in bank departments **Methodological support** to both **systemically important financial institutions** and **supervisory entities**

Solution

Advanced **software solutions** for **modelling, forecasting, calculating** metrics and **integrating** risks, all on cloud and distributed in Software-as-a-Service (**SaaS**)

KEEP IN TOUCH

iason

Company Profile

iason is an international firm that consults Financial Institutions on Risk Management. **iason** integrates deep industry knowledge with specialised expertise in Market, Liquidity, Funding, Credit and Counterparty Risk, in Organisational Set-Up and in Strategic Planning.

Margherita Ranieri

This is an **iason creation**.
The ideas and the model frameworks described in this presentation are the fruit of the intellectual efforts and of the skills of the people working in iason. You may not reproduce or transmit any part of this document in any form or by any means, electronic or mechanical, including photocopying and recording, for any purpose without the express written permission of iason.

www.iasonltd.com

Copyright © 2026 - All rights reserved