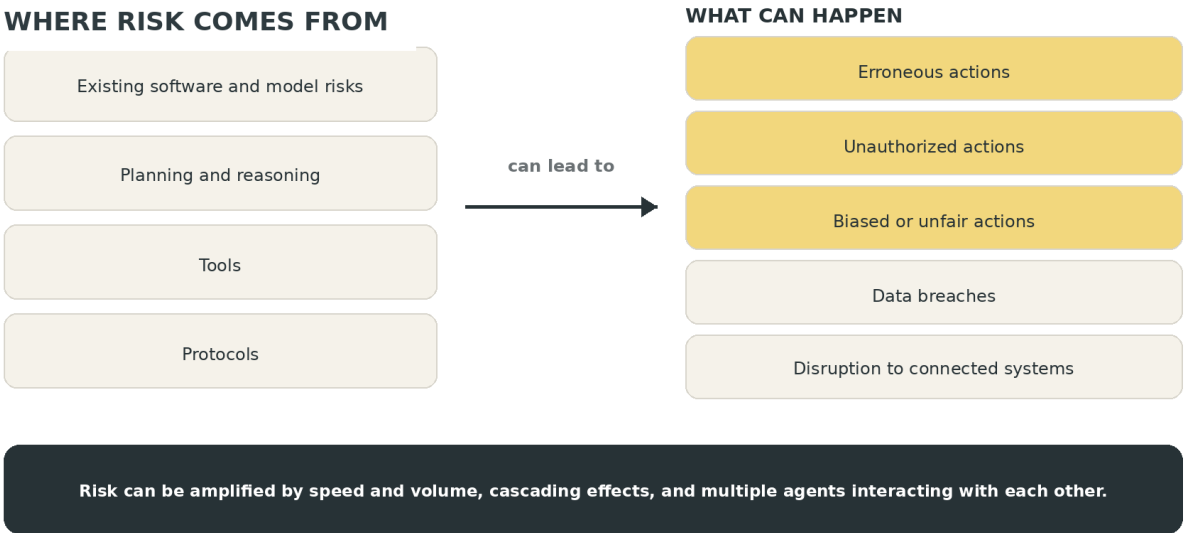


AI agent risks

This guide explains where AI agent risk comes from, the outcomes that matter, and the factors that can increase the impact of a failure.

AI agents inherit familiar software and generative AI risks. They can produce inaccurate output, reflect bias, leak data, or be manipulated. What changes is that agents add planning, tools, and connections that allow them to act. A simple model error can therefore become a real action with real-world consequences.

For a first review, it helps to separate three questions: (1) where risk comes from; (2) what harmful outcome could follow; and (3) what could amplify the impact. That structure is used below.



Sources of risk

Source of risk	What it means in practice
Existing software and model risks	Agents still inherit software vulnerabilities and model risks such as inaccurate output, bias, data leakage, and prompt injection.
Planning and reasoning	The agent can misunderstand the goal, create a poor plan, or drift away from the original task as it works through several steps.
Tools	The agent can call the wrong tool, use the right tool with the wrong input, or take an action outside the intended scope.
Protocols	A connection to a tool or another agent can be misconfigured, untrusted, or compromised, which can expose data or redirect the workflow.

Five possible AI agent risks

Erroneous actions	The agent takes the wrong action.
Unauthorized actions	The agent acts outside its permitted scope or skips a required approval.
Biased or unfair actions	The agent produces actions that lead to unfair treatment or outcomes.
Data breaches	The agent exposes or wrongly changes sensitive information.
Disruption to connected systems	The agent damages or disrupts a system it can reach.

Why agent risk can spread quickly

The same underlying error can have a larger effect when an agent acts quickly, carries an error into later steps, or works with other agents.

Speed and volume: Agents can take many actions at machine speed. A problem can repeat before a person notices.

Cascading effects: An error in one step can become the input for the next step and grow through the workflow.

Multi-agent complexity: When several agents share information or hand work to each other, miscoordination and unexpected behavior become harder to predict.

Four governance questions to start with

If your organization is asking these questions, Ortaire can help you turn them into a practical governance approach.

- 1. What can the agent access?** Which data, files, systems, or other agents can it reach?
- 2. What can the agent do?** Can it only read and draft? Or can it also update, send, approve, delete, or trigger another process?
- 3. How much can it decide on its own?** Does it follow detailed instructions, require approval at key steps, or choose most steps independently?
- 4. What happens if it is wrong?** Identify the realistic consequence and whether speed, volume, or connected agents could make the impact larger.

Contact us: www.ortaire.com/contact or email hello@ortaire.com

Example: internal Legal Q&A

Source of risk: The agent relies on an outdated policy or follows a poor plan when a question falls outside its scope.

Possible outcome: The business receives an incorrect answer or the request is routed incorrectly.

What could amplify it: High question volume could repeat the same error across many requests before it is detected.

Practical response: Limit the agent to approved sources and defined topics, route uncertainty to Legal, and monitor corrections and repeated failure patterns.

Key takeaway: Agent risk is not only about whether the model gives a wrong answer. It also depends on what the agent can access, what it can do, how independently it can act, and how quickly a mistake can spread.

Ortaire provides non-legal consulting that helps organizations adopt, govern, configure, and implement AI in practice. We are not a law firm and do not provide legal advice, representation, or compliance determinations. Working with Ortaire does not create an attorney-client relationship or privilege. Organizations should consult their own counsel for legal advice.