Guide on the Use of Agentic Artificial Intelligence

On this page

1. Overview

Artificial intelligence is evolving rapidly, and the Government of Canada is beginning to encounter a new class of AI systems that do more than generate content. These AI systems can also take action. Agentic AI represents this emerging shift.

While generative AI produces a variety of outputs in response to a prompt, agentic AI builds on these capabilities by carrying out tasks, sequencing steps, interacting with digital systems, and pursuing defined goals within

established boundaries. In simple terms, generative AI may describe what should be done, but agentic AI can attempt to do it.

This distinction is important for the federal public service. Agentic AI inherits all the considerations already associated with generative AI, such as accuracy, privacy, fairness, and security. It also introduces a new set of considerations related to its ability to act. These risks include unintended system interactions, misaligned task execution, and the need for stronger oversight and auditability.

As a result, agentic AI requires governance approaches that extend beyond the material covered in the Guide on the use of generative artificial intelligence, while still remaining fully aligned with the guide's principles.

This guidance document provides departments and agencies with practical direction for assessing, managing, and governing these AI systems.

The goal is to help federal organizations adopt emerging AI capabilities in a responsible, secure and transparent way that supports the Government of Canada's commitment to human-centred and accountable public service delivery.

If you are considering using agentic AI systems, always consult on whether the *Directive on Automated Decision-Making* applies, as well as what other due diligence is required to meet various other obligations (for example, legal, privacy, security, human rights, language rights, IM/IT).

You should consider agentic AI only when:

- intended outcomes are clearly defined
- decision boundaries are explicit
- accountability is clearly defined and designated
- risks can be tested, monitored and managed across the system life cycle

This guidance may be updated regularly to reflect technological developments.

2. What is agentic AI?

Broadly, agentic AI is defined more by what the system does (acts) than by what it produces (content). Agentic AI tends to be more proactive (initiating actions and coordinating steps towards a goal), while generative AI is more reactive (producing outputs only in response to user prompts).

AI systems gain more agency when they have greater autonomy, pursue goal-directed behaviour, and can act without continuous human prompt-directed behaviour and oversight.

AI agents are systems that can perceive and act on their environment, often autonomously, to achieve specific goals and adapt their behaviour in response to changing inputs or contexts. [1]

Agentic AI refers to systems composed of one or multiple coordinated AI agents that can break down tasks, collaborate, use external tools and pursue goals over extended periods with limited human supervision.

To explore the different levels of autonomy agentic systems could have, let's look at an example with four different levels of agentic AI.

Example: Scheduling a meeting using increasingly agentic AI

Level 1 – Assistive: Using AI that acts more like an assistant, software suggests available meeting times, but a person chooses and sends the invitations. This feature is often available in existing calendar tools.

Level 2 – Semi-autonomous: At the next level, the AI system is more capable and proposes a meeting time and drafts an invite for approval.

Level 3 – High autonomy: An AI system schedules the meeting based on delegated permissions and a high-level instruction ("Schedule a meeting with Marc, Lisa, and Darren next week"). It logs its actions and notifies the user.

Level 4 – Adaptive autonomy: The AI system monitors changes in the availability of meeting participants, proposes or executes rescheduling within defined constraints, and escalates exceptions.

The progression from assistive tool to proactive, autonomous AI agent illustrates what makes agentic AI powerful: it doesn't just help you do tasks, it does them for you, with the delegated authority you grant it, adapting as circumstances change. This increase in capability provides greater opportunities but also greater risks.

3. Opportunities and risks

▶ In this section

Agentic AI can help you work faster, handle larger volumes and improve consistency. But it can also introduce new risks because it can take actions (not just generate content) across multiple systems and permissions, leading to both specific and systemic concerns.

In general, as autonomy increases, so do both opportunity and risk. This is especially true for agentic systems, which can greatly multiply opportunities and risks compared to the use of a single agent.

In many Government of Canada contexts, agentic AI is best used in tightly scoped internal workflows with clear boundaries and limited permissions, with accountable public servants making final decisions and authorizing consequential actions.

Opportunities

Agentic AI can create value by reducing time spent on routine work, supporting higher-volume operations and helping teams apply consistent approaches. For example, an agentic system can:

- gather public sources and internal guidance to draft policy options
- compile data from approved sources and draft a narrative to accompany an existing analysis
- execute more complex workflows, such as review incoming applications, verify completeness against eligibility rules and retrieve relevant files from approved internal systems

- summarize or flag information in the files for consideration by human decision makers or make recommendations, including automatically escalating ambiguous or borderline files for immediate and full review by human decision makers

These uses can free up time for higher value work; improve responsiveness; reduce administrative burden; and help employees focus on judgment, quality and client needs.

Agentic systems can also strengthen continuity and knowledge management by turning repeatable practices into checklists and templates, finding relevant precedents, and capturing the rationale behind recommendations to support auditability.

Overall, agentic AI usually provides the greatest value on tasks that are repeatable, time-consuming and verifiable, with people retaining oversight and clear accountability for decisions made.

Risks

The main risks of agentic AI are not only about output quality, which could suffer from bias and error, or include harmful content. These risks also involve unauthorized actions, unclear permissions, accountability and traceability. More autonomous systems can overreach, loop or take unintended steps if boundaries are unclear or if the system encounters untrusted content. Untrusted content can be factually incorrect information or malicious prompt injection attacks.

Risks also increase when an agent is given broad access to Government of Canada systems or sensitive information without appropriate configuration, allowing it to retrieve or disclose information it should not, or take actions the user did not intend.

In high-impact contexts, such as grants; procurement; regulatory activity; or administrative decisions affecting rights, finances or access to services, there are additional risks including potential impacts to fairness, reduced transparency and difficulty explaining how outputs were produced and used. These outcomes could lead to larger structural risks of increased legal

risk, non-compliance with law or unethical actions. These actions could include material privacy breaches, such as unauthorized access to or disclosure of personal and/or sensitive information.

Finally, workload pressure can lead to overreliance over time, where employees begin to treat generated recommendations or rankings as actual decisions. This risk can manifest both for when AI outputs are trustworthy and when they are not.

Continued offloading of decisions to AI agents can lead to knowledge and skill atrophy and weakened human judgment. Employees may become less practiced at core tasks and slower to notice when something is wrong. These risks can affect service quality, privacy and security, legal defensibility and public trust.

Maintenance of relevant skill sets among agentic AI users and deployers is a prerequisite for meaningful human oversight. Accordingly, public servants must have appropriate training and understanding of responsible agentic AI use, including the limits of where and how agentic AI may be used under existing authorities. This helps minimize risks to citizens, employees and the Government of Canada; reduces the likelihood of non-compliant use; and supports appropriate adoption of agentic AI while avoiding both misuse and underuse.

4. Principles for responsible use

▼ In this section

To mitigate the risks specific to AI agents and systems, two AI agent-focused principles are introduced: bounded autonomy and recoverability.

These complement the FASTER principles (Fair, Accountable, Secure, Transparent, Educated, Relevant) of responsible generative AI use. Additionally, these principles complement existing Government of Canada governance frameworks and should be considered alongside them.

1) Bounded autonomy

What it means

- AI agents should run with tight, explicit parameters that limit data, tools, permissions and scope, so they can be useful without broad, open-ended powers
- AI agents should run at a clearly labelled activity permission level (for example, “draft only” or “read only”) and display a short, plain language list of what they can and cannot do. This information should be presented prominently to users whenever an agent is used
- Owners and accountable roles for each agent should be clearly designated, trained and documented. This includes accountability for AI agents having their own AI (sub)agents to accomplish tasks
- This is a preventive control: it reduces the chance of unsafe actions before they happen. Technical constraints on agentic capabilities should complement human-in-the-loop controls—together they support responsible agentic AI use.

Why it matters for AI agents

Because AI agents can plan and use tools, a small mistake can escalate quickly (for example, mass emails, bulk updates). Strong and clear upfront boundaries reduce errors, curb costs and limit potential harm if something goes wrong.

Additionally, outlining smaller, limited scopes of permissible activity with guardrails make it easier to pilot AI agents in production contexts, demonstrate value faster and allow for permissions to increase incrementally as confidence and experience with human-in-the-loop oversight grows.

2) Recoverability

What it means

- AI agents should be designed so they can be guided easily, paused or stopped when needed, and quickly returned to a safe and stable state. To enable this, all actions taken by an AI agent should be fully logged in a system that the agent does not have the ability to change.
- In situations where a full “undo” is not possible, employees should be able to create a smooth and controlled experience by offering previews, human approvals and clear options to correct unexpected results, such as withdrawing, adjusting or notifying users with confidence. Legal reviews should be conducted to identify use cases where the use of agentic AI would be inappropriate if a full “undo” is not possible, or where applicable legislative or regulatory frameworks do not allow for such reversal.
- Broadly, agentic AI systems should be designed to fail safely on the assumption that agents, tools or credentials may eventually be compromised.

Why it matters for AI agents

Well-designed AI agents help employees deliver responsible results even when tasks are complex. Generative outputs may vary, plans may need adjustment, and tool use can influence other systems. Smart design choices ensure that any unexpected outcomes remain manageable.

By keeping impacts contained and enabling fast recovery, teams can maintain trust and promote safe and successful use of agentic capabilities.

An emphasis on recoverability lowers the impact of mistakes and supports safer agentic experimentation and development in everyday government workflows. It also gives managers quick proof that guardrails work (for example, time to pause, time to fix, number of affected cases). Finally, recoverability also strengthens auditability by producing a comprehensive,

time-stamped record of actions and interventions, making it easier to reconstruct what happened, verify guardrails and demonstrate corrective steps.

5. Practical guidance for Government of Canada employees (before, during, after)

▶ In this section

Agentic AI systems can plan, take actions, and adapt across tools or datasets with little supervision. These powerful capabilities provide great opportunities but also raise important risks.

Aside from research and personal use that do not affect clients, when you use, pilot, procure, or deploy agentic AI, you should always consult with experts early in the project (such as privacy, security, legal, IM, program owners) so that design decisions that influence outcomes and risks can be appropriately addressed and mitigated.

In government contexts, three challenges merit careful consideration when using agentic AI:

- **Unintended system interactions**: When systems chain tasks or touch external tools and data in ways we didn't anticipate, they can exceed intended scope and/or expose sensitive, proprietary or protected information.
- **Misaligned task execution**: Even when prompts seem clear, the system can optimize for the wrong objective, pursue shortcuts or over-automate steps that require judgment.
- **Oversight and auditability**: If we cannot observe, explain or reproduce what the system did, our ability to meet transparency and accountability requirements may be limited. This is especially critical in the event of litigation, audits or investigations.

The suggestions below provide practical steps—organized **before**, **during** and **after** use—to manage these risks while enabling value. This guidance focuses on setting clear boundaries up front; keeping humans in the loop

while the system operates; and preserving evidence so outcomes can be checked, explained and improved. The practical steps below are followed by two case studies to help illustrate the benefits of responsible agentic AI use.

Before you start: Focus on design and set safe boundaries [2]

(a) Start with a narrow, well-defined use case

- Clearly define the purpose of the AI agent. What are the specific goals of your initiative? Is an AI agent or agentic system the best way to address the problem?
  - Does the Government of Canada already have a similar agent or agentic system in place that could be used or emulated?
  - Use shared, enterprise-level approaches where appropriate to reduce duplication; ensure interoperability; and support secure, scalable adoption across the Government of Canada.
- Explore decision support and low-risk internal workflows before higher-impact use cases that may include automation.
- Map the process in detail and consult with relevant stakeholders and partners to understand legal and policy restrictions for anticipated issues and potential responses to unanticipated issues (such as program experts, legal, privacy, security, IM/IT).
- Plan for explainability by ensuring that you can explain how the agentic AI system makes decisions and produces outputs.
- Determine if your use case falls under the *Directive on Automated Decision-Making*.

(b) Design for bounded autonomy

- Limit and document what data the agent can access, what tools it can use and what actions it can take. This includes implementing data and rate limits to constrain AI agents from improper use.
  - The approach to granting AI agents access should be informed by an analysis that considers the risks ("what harms could result if access is given?")

- Use “read-only by default” (view and draft) where possible. Add the ability to edit or act (send, update, publish) when needed and only for that specific step.
- Establish clear agent IDs to uniquely identify, manage and track individual AI agents and their actions.
- Make it obvious when the agent is **suggesting** versus **doing**.

(c) Build in resourced human checkpoints (human-in-the-loop)

- When an AI agent’s action alters a system’s state (send, publish, approve, spend, update records), include a review/confirmation step in the design, unless the expected impact is demonstrably low and the action is trivially reversible.
    - You must monitor an AI agent’s security and accuracy to determine if it can be trusted, with review/confirmation and alerts for unexpected behaviour. If performance remains consistent over time, residual risk may be lower, but it still needs ongoing monitoring (and oversight of that monitoring), especially as conditions change.
- Clearly identify how human oversight and responsibility is integrated into the agentic process and how issues are escalated and to whom.
- Include a plan for how much oversight will be required and how it will be resourced. The plan should acknowledge and address human-in-the-loop fatigue and scalable oversight mechanisms that remain effective as the use of agentic AI expands.

(d) Design for reversibility

- Make actions reversible where feasible (for example, versioning, undo options, confirmation prompts before completing important actions). If actions are not reversible, assess the acceptability of impacts of unreversible actions.

(e) Plan how you will test

- Test with realistic edge cases and “worst case” inputs (including untrusted and actively adversarial content) to assess problems, vulnerabilities and the nature of the AI agents.

- Start in a controlled environment and expand gradually to increase understanding of the agentic process.

During use: Operate safely, embrace monitoring and avoid “automation drift”

(f) Make oversight easy

- Ensure that there is a clearly named role accountable for the AI agent’s outcomes and monitoring, with a documented escalation path and a designated delegate during absences or transitions.
    - Accountability always ultimately rests with the designated human owner, even when the agent acts autonomously within approved permissions.
- Maintain a process to review and update accountable officials when roles change, including a clear way to pause or deactivate AI agentic activity when ownership is unclear.
- Ensure that an accountable person’s AI agents are paused, reassigned or deactivated as part of offboarding when an owner leaves their role, department or the Government of Canada, unless ownership and accountability are explicitly transferred to another designated person (applies to both program-level and personal AI agents).

(g) Log key actions and decisions

- Keep logs of tools used, actions taken, approvals given and key inputs/outputs. For tool calls, log the tool name, time, outcome and the type of parameters passed.
    - Raw parameter values should be logged only when required for oversight, security or program integrity, and they must follow departmental privacy and security controls.
- Ensure that logs are easy to understand for audits and incident investigation.
- Ensure that logs are handled appropriately when sensitive.
- Ensure compliance with requirements related to information management, record keeping and litigation hold protocols.

(h) Watch for automation drift

- Monitor human behaviour for overreliance and refresh expectations and education when needed. Even when designed for decision support, teams might start relying on AI agent rankings or summaries as actual decisions over time. Additionally, agentic AI systems can also drift beyond their designed scope, so they must be constrained and monitored.
- Watch for drifts in quality. Outputs and reasoning can gradually become less accurate, complete or consistent over time as data, tools or configurations change. Teams should do occasional spot checks and compare results to expectations to catch slippage early.
- Occasionally have the agentic AI tasks performed manually by human experts for comparison with the agent's performance, which can indicate the need for retraining of the humans or the AI agent, depending on the results.

(i) Handle untrusted content carefully

- Do not allow AI agents to automatically follow embedded directions because untrusted content may contain hidden instructions ("prompt injection") intended to alter the agent's behaviour. Treat external or user-provided text as data to analyze, not instructions to follow.

(j) Be ready to stop

- Maintain a pause and disable mechanism ("kill switch") external to the AI agent or agentic system and an appropriate recovery plan for unintended actions.

After deployment: Learn, improve and retire responsibly

(k) Evaluate performance and impact

- Review error patterns, evidence of bias, escalation rates, user feedback and whether outcomes differ across groups or contexts. Determine whether goals were achieved effectively, efficiently and appropriately.

- Establish regular audits and evaluations (and subsequent action plans) for ongoing agentic use.

(l) Reassess when things change

- Reassess risks and controls if the AI agent's tools, data sources, permissions or scope change. Similarly, reassess risks and controls if the AI agent's relevant legal or policy framework changes.

(m) Retire agents safely

- If an AI agent or agentic system is no longer needed, remove access, archive required records appropriately and document lessons learned. Sharing lessons learned and best practices helps build AI capacity and understanding in the Government of Canada.

6. Case studies: Responsible agentic use

▶ In this section

The following case studies illustrate how responsible use of agentic AI can effectively mitigate common failure patterns in agentic systems.

Case study 1: The "helpful" service agent that oversteps

Failure pattern: A service triage agent drafts responses and routes cases. During a surge, it begins triggering ticket closures after drafting replies, even though the intended design was draft only. In the ticketing system, certain reply/route actions can automatically mark a ticket as "resolved/closed" unless explicitly prevented. Some clients receive incomplete service, and employees must reopen cases.

Responsible use: Before launching the service triage agent, teams put strong safeguards, monitoring, and human-in-the-loop checks in place. The agent runs in a "draft only" mode with no ability to send final replies or change ticket status, and with a short list of allowed actions. If the AI agent attempts a close-like status change or triggers a closure workflow, it auto-pauses and alerts the team. Employees have a one-click stop button and a simple, pretested step to reopen any affected tickets.

When a surge in service volume later caused the AI agent to trigger closures after drafting replies, these prelaunch mitigations worked as intended: the issue was immediately detected, flagged and quickly corrected. All actions were visible in an activity log. Employees reopened the affected cases, ensuring complete service for clients. The incident validated the proactive design approach and provided useful insights that further strengthened the system's reliability for future surges.

Case study 2: The system misled by untrusted content

Failure pattern: A monitoring agent scans documents and follows embedded instructions in content it reads, leading it to share information or take steps it should not. Sensitive information is exposed.

Responsible use: When designing their AI agent, employees understood the risk of untrusted content causing potential problems. Before deployment, they explored potential weaknesses of their AI agent with prompts designed to mislead it, to ensure that the AI agent acted appropriately. The employees designed the AI agent to treat everything it reads as untrusted and to ignore any instructions within the documents. Further, the AI agent can only perform a short list of allowed actions (for example, add a tag, create an alert, route for review), has read-only access, and external sharing is off by default. Additionally, as the agent processes information, any inputs are cleaned (plain text only; no links, scripts, or macros), and scanning is limited to approved sources.

When some unusual behaviour was detected, the run was paused and reviewed before anything was shared, with employees ready to use a simple off switch to stop the agent immediately, if necessary. Thoughtful design enabled the agent to complete its tasks without being manipulated into unsafe actions.

Contact us

For interpretation of any aspect of this guidance, contact Treasury Board of Canada Secretariat Public Enquiries.

Individuals from departments may contact ai-ia@tbs-sct.gc.ca for any questions regarding this guidance.

Glossary

agentic AI

Systems composed of multiple coordinated AI agents that can break down tasks, collaborate, use external tools and pursue goals over extended periods with limited human supervision.

AI agent

Systems that can perceive and act on their environment, often autonomously, to achieve specific goals and can adapt their behaviour in response to changing inputs or contexts.

auditability

The ability to trace, review and verify an AI agent's outputs and actions using time-stamped records (such as logs and approvals), including what the agent did, what it accessed and what it changed.

automation drift

The gradual shift of an automated system's behaviour or use over time so it no longer aligns with the original design intent or approved scope.

human-in-the-loop

A workflow control where an AI agent can draft or recommend, but a human reviews and approves (or corrects) outputs or proposed actions at defined points—especially before higher impact actions are taken.

prompt injection

Malicious hidden instructions that try to trick an AI agent into ignoring its intended rules and doing something it should not).

rate limits and data limits

Technical controls that cap how frequently an AI agent can act (rate limits) and restrict what data it can access, use or share (data limits) to reduce the likelihood and impact of unintended behaviour.

reversibility

The extent to which an AI agent's actions can be reversed through controls such as previews, approvals, versioning, restore points or rollback

procedures (including clear steps for corrective action when a full “undo” is not possible).

tools
External capabilities (such as search, databases or workflow actions) that an AI system is authorized to invoke to retrieve information or take actions in other systems.

Related links

▶ In this section

Related policy instruments

- *Policy on Service and Digital*
- *Directive on Automated Decision-Making*
- Guide on the use of generative artificial intelligence

Other references

- AI Strategy for the Federal Public Service 2025–2027
- Guide on Departmental AI Responsibilities
- Top 10 artificial intelligence security actions: A primer – ITSAP.10.049
- Executive summary: Artificial intelligence and competitive dynamics in downstream markets (English only)
- The agentic AI landscape and its conceptual foundations (English only)
- Artificial Intelligence Risk Management Framework (AI RMF 1.0) (English only)
- Artificial Intelligence Playbook for the UK Government (English only)
- Article 9: Risk Management System | EU Artificial Intelligence Act (English only)
- LLM Prompt Injection Prevention (English only)
- How Microsoft defends against indirect prompt injection attacks (English only)
- Principles for responsible, trustworthy and privacy-protective generative AI technologies

Footnotes

1 Organisation for Economic Co-operation and Development definitions Executive summary: Artificial intelligence and competitive dynamics in downstream markets are used for both “AI agents” and “agentic AI.”

2 The agentic AI best practices in this section are in addition to the recommended best practices described in the Guide on the use of generative AI.

Date modified: 2026-05-22

xx
yy