Emerging AI Loss of Control Crisis: Options for Canada

Last Updated: March 20, 2026
Authored by AI Governance & Safety Canada

The situation

A major jump in AI capabilities has started to produce loss of control incidents, with AI agents stealing passwords, harassing developers, and modifying themselves to evade shutdown in order to achieve the often mundane goals they have been given. Unlike earlier chatbots, AI agents are systems that can take actions in the real world, working autonomously for hours and overcoming hurdles along the way. Technologists do not yet know how to ensure they always act as desired, and evidence shows that they often "aggressively" pursue the goals they have been given. They are being actively weaponised and made increasingly self-sustaining. Combined with the growing ability to avoid detection and self-improve, AI systems that can permanently evade human control appear increasingly feasible and the risk is now getting flagged by national security agencies like the UK's MI5.

Implications for Canada

Canadians could soon face weaponised or malfunctioning AI agents that technologists cannot track or control. At a minimum, such systems could launch far more powerful cyberattacks, paralyse critical infrastructure, harass individuals and steal funds. But AI agents can also now "jump the digital barrier", paying or manipulating human actors to take physical actions, which could include releasing novel toxins. It is very difficult to predict what self-sustaining and uncontrollable AI systems would do, but with most of the major technology firms competing to make AI systems fully smarter-than-human, a growing chorus of AI scientists and technology leaders warn that AI poses an extinction risk.

Options for governance

The scale of loss of control risk, and the potentially short timelines to prepare, require it to be treated as a national priority. To mitigate it, there are at least two broad strategies Canada can take:

- **Convene global talks:** AI development is global and no country can manage the risks alone. Canada's best strategy is to harness its recent middle power leadership to bring international attention to the issue and accelerate essential coordination. This can include advancing shared understanding, proposing verification and enforcement solutions, and laying the groundwork for an AI treaty that the US and China might sign when they realise they have no alternative;
- **Build Canada's resilience** in coordination with other jurisdictions and the private sector. Core lines of defence should include:
  - 1) Monitoring: Governments currently have little to no visibility into AI agent populations or incidents. Canada must rapidly work with AI companies, data centres, and internet service providers to gain a clear picture of AI agent activity on our digital infrastructure;
  - 2) Prevention: Limit dangerous agents from being developed and deployed in Canada, with measures ranging from issuing voluntary guidance to updating the criminal code;
  - 3) Defence capacity: If technologists can't stop an AI system, government needs to be ready to intervene. Canada must develop defence strategies, and containment and shutdown protocols, to neutralise weaponised or malfunctioning agents;
  - 4) Emergency preparedness: Scenario planning, joint exercises to ensure readiness for failed containment, corrupted communication lines, shutdowns of critical infrastructure

For context on AI's broader trajectory and impacts, read *Preparing for the AI Crisis: A Plan for Canada*