DATARISK CANADA / AI SECURITY

# AGENTIC AI CYBERSAFETY CHEAT SHEET

A free self-assessment based on the six fundamental AI cybersecurity principles of the Datarisk AI Security Hexad

www.AI-Security.ca | Classification : PUBLIC | Claudiu Popa

**Why use this Cheat Sheet?**
An AI agent can read sensitive data, use tools, connect to services and impact systems. Unplanned access, unapproved actions or missing safeguards can turn an experiment into a serious incident. **Share this completed form with your IT team before connecting an AI agent to real data or systems.**

**AGENT / PROJECT** _____ **OWNER** _____

**ENVIRONMENT** ☐Test ☐Pilot ☐Production **DATE** _____

Mark YES with evidence, or GAP for unresolved issues. IT/Security must validate and approve before release.

| THE SIX FUNDAMENTAL AI CYBERSECURITY CONTROLS | YES | GAP |
|---|---|---|
| **01 Identify the Agent and the Accounts it Uses**<br>Does your agent have a distinct identity, named human owner, approved purpose and traceable credentials?<br>*Be ready to provide proof. For example: name the owner, purpose, agent ID and access account.* | ☐ | ☐ |
| **02 Limit Access Privileges**<br>Are the agent's privileges, access destinations, runtime and (token) spending limited to what the (approved) task requires?<br>*Evidence: permission list, allowed tools and expiry/quotas.* | ☐ | ☐ |
| **03 Contain and Constrain**<br>The agent is isolated; network access and sensitive systems are blocked unless explicitly authorized.<br>*Evidence: sandbox settings, network rules and test results. These can rely on existing network controls.* | ☐ | ☐ |
| **04 Verify and Approve**<br>Inputs and actions are checked against policy and expected data; risky or irreversible actions require human approval first.<br>*Evidence: action policy, approval thresholds and sample approvals.* | ☐ | ☐ |
| **05 Observe and Detect**<br>Tool calls, data access, changes, errors and approvals are traceable in protected logs and live monitoring with proper alerts.<br>*Evidence: audit trail and an alert/detection test.* | ☐ | ☐ |
| **06 Terminate and Recover**<br>The agent can be stopped, credentials revoked and access isolated; recoverable changes can be restored.<br>*Evidence: tested stop, revocation and recovery procedure.* | ☐ | ☐ |

## AFTER THE SIX CHECKS / FURTHER REVIEW AS NEEDED

- **Privacy & data protection**
  PIA, lawful use, sensitive data, retention, transfers.
- **Adversarial security testing**
  Threat model, prompt injection, misuse, exfiltration.
- **AI governance & impact**
  Safety, reliability, bias/human rights, human oversight.
- **Suppliers, standards & law**
  Vendor/API review; ISO 27001/42001, NIST AI RMF; sector rules.

## HAND OFF TO IT / SECURITY · REQUIRED BEFORE RELEASE

Provide this page and links/IDs for configurations, logs, tests and unresolved gaps. Never include passwords or keys.

**EVIDENCE / TICKET** _____ **REVIEWER** _____

Reviewer decision: ☐Approved for scope ☐Changes required **DATE** _____

## INDEPENDENT SECURITY VERIFICATION

Datarisk Canada | Verify™ AI Security Assessment · Request a review: AI-Security.ca | VerifyNow.ca

This free self-check does not constitute independent verification, certification or approval to deploy.