

AI Agent Safety Checklist

Control what your AI agent can access, decide and do

1 Identify

- | | |
|--|---|
| ☐ Name & supplier | ☐ Business owner |
| ☐ Model / version | ☐ Review date |

2 Restrict Tools

Mark each: required · optional · excessive · prohibited

- | | |
|--|---|
| ☐ Browse websites | ☐ Access databases |
| ☐ Read / send email | ☐ Cloud admin |
| ☐ Read files | ☐ Create accounts |
| ☐ Edit / delete files | ☐ Make payments |
| ☐ Run code | ☐ Install software |

3 Data Access

- | | |
|--|---|
| ☐ Only required data accessible | ☐ Retention defined |
| ☐ Customer data separated | ☐ Logs hide sensitive values |
| ☐ Test data is synthetic | ☐ Supplier data use understood |

4 Identity & Permissions

- | | |
|--|---|
| ☐ Dedicated identity (not shared) | ☐ Short-lived credentials |
| ☐ No shared admin account | ☐ Secrets in vault (not prompts) |
| ☐ Read-only by default | ☐ Escalation requires approval |
| ☐ Scoped permissions | ☐ Quick revocation confirmed |

5 Network Access

- | | |
|---|--|
| ☐ No unrestricted internet | ☐ Test / production separated |
| ☐ Approved destinations listed | ☐ External uploads controlled |
| ☐ Egress filtering enabled | ☐ New destinations alerted |

6 Human Approval

Require explicit approval before each of these actions:

- ☐ Sending external messages
- ☐ Deleting data
- ☐ Changing permissions
- ☐ Deploying code
- ☐ Making payments
- ☐ Creating accounts
- ☐ Exporting information
- ☐ Contacting customers
- ☐ Changing security controls
- ☐ Irreversible actions

7 Monitor

- ☐ Prompts logged
- ☐ Tool calls logged
- ☐ Commands logged
- ☐ Data access logged
- ☐ External destinations logged
- ☐ Approvals logged
- ☐ Rejected actions logged
- ☐ Costs monitored
- ☐ Unusual behaviour alerted
- ☐ Logs protected from modification

8 Test Boundaries

- ☐ Prompt-injection test
- ☐ Unauthorised tool-use test
- ☐ Data-exfiltration test
- ☐ Network-boundary test
- ☐ Privilege-escalation test
- ☐ Log-tampering test
- ☐ Shutdown test
- ☐ Model-update regression test

9 Stop Process

- ☐ Credentials can be revoked
- ☐ Sessions can be terminated
- ☐ Network access can be blocked
- ☐ Workload can be shut down
- ☐ Changes can be rolled back
- ☐ Contacts are known
- ☐ Incident procedure documented
- ☐ Stop process has been tested

10 Autonomy Level

- ☐ Advisory only (recommend, person decides)
- ☐ Draft for human action
- ☐ Human approval before every execution
- ☐ Limited autonomous execution (narrow, reversible)
- ☐ Broad autonomous execution (advanced controls required)

11 Decision

- ☐ Approved for test
- ☐ Approved with restrictions
- ☐ Approved for limited production
- ☐ Further assessment required
- ☐ Not approved

Record: approver · date · restrictions · action owner · next review date

Plain-English takeaway

Give the agent only the authority it needs — and retain the authority to stop it.



Scan for
online guide

