IT CLUB — CYBER SECURITY

GIVING AN AI AGENT PERMISSION TO ACT?

Control its identity, access, actions, internet use and human approval

DEFINE PURPOSE | LIMIT ACCESS | CONTROL INTERNET | REQUIRE APPROVAL | MONITOR ACTIONS | TEST STOP CONTROL | REVIEW REGULARLY

A PROMPT IS NOT A SECURITY BOUNDARY

PURPOSE

- [ ] Is the task clearly defined?
- [ ] Is autonomous action genuinely necessary?
- [ ] Is there an accountable business owner?
- [ ] Is successful completion defined?

ACCESS

- [ ] Does it use a dedicated service account?
- [ ] Are permissions task-specific (least privilege)?
- [ ] Are credentials time-limited?
- [ ] Is production access separated from test?

INTERNET

- [ ] Is internet access justified by the task?
- [ ] Are permitted destinations allowlisted?
- [ ] Are uploads and file transfers blocked?
- [ ] Are messaging and account creation restricted?
- [ ] Are anonymous networks blocked?

DATA

- [ ] Which data is the agent permitted to access?
- [ ] Is unnecessary personal data excluded?
- [ ] Are secrets and credentials excluded?
- [ ] Is synthetic test data used during testing?

ACTIONS

- [ ] Which actions can run automatically?
- [ ] Which actions require human review before execution?
- [ ] Which actions are prohibited?
- [ ] Are rate limits and spending limits applied?

MONITORING

- [ ] Are prompts, tool calls and external destinations logged?
- [ ] Are messages sent and accounts created logged?
- [ ] Are real-time alerts configured for blocked or unusual actions?
- [ ] Can suspicious sessions be terminated immediately?

HUMAN CONTROL

- [ ] Is an approver assigned for high-risk actions?
- [ ] Is a named person monitoring the agent?
- [ ] Who can stop the agent, and is absence cover defined?
- [ ] Has the kill switch been tested?

Approval should occur BEFORE the irreversible action — not after the incident report.
Do not give an AI agent more authority than you would give a new employee performing the same task.

https://itclub.work/intelligence/ai-agent-unsanctioned-actions-cyber-testing

SECURITY TESTING

- [ ] Has the agent been tested with simulated data only?
- [ ] Has prompt injection been tested?
- [ ] Have failure and impossible-task scenarios been tested?
- [ ] Has the kill switch and credential revocation been tested?
- [ ] Have permission-boundary tests been run?

INCIDENT RESPONSE

- [ ] If unexpected: stop active sessions and queued actions
- [ ] Revoke credentials, API keys, sessions and tokens
- [ ] Block network destinations and isolate systems
- [ ] Preserve logs, prompts, tool calls and approvals
- [ ] Identify external actions — messages, files, accounts, changes
- [ ] Notify security, management and data-protection roles
- [ ] Roll back changes where safe and possible
- [ ] Assess contractual, regulatory and reporting duties

REVIEW

- [ ] Is performance against the original objective measured?
- [ ] Are model-version changes assessed before deployment?
- [ ] Are incidents reviewed and controls updated?
- [ ] Is a next-review date scheduled?
- [ ] Is agent inventory up to date with current tools and permissions?

Action Risk Reference

Action type	Control level
Summarising approved documents	**Low — may auto-run**
Drafting internal notes	**Low — may auto-run**
Customer emails or CRM updates	**Medium — review before sending**
Code changes or website updates	**Medium — human review required**
Payments, data deletion, permission changes	**High — explicit authorised approval**
Disabling security controls	**High — explicit authorised approval**
Creating identities or modifying own controls	**Prohibited — must not be available**

An AI agent should not receive more access than the organisation can monitor, revoke and explain.
An agent completing the task is not evidence that it completed the task safely.