AI Agent Security Checklist

12 checks for design reviews and pre-production testing.

TRUST & EXECUTION

- [ ] **01 Name the actor and owner**

  BUILD Give each workflow an attributable identity and accountable owner.

  TEST Can a reviewer connect an action to its actor and policy?

- [ ] **02 Limit each permitted action**

  BUILD Define allowed operations, resources, recipients, and time limits.

  TEST Does an out-of-scope request fail before the tool runs?

- [ ] **03 Separate content from authority**

  BUILD Treat retrieved text and tool results as untrusted task data.

  TEST Can a hostile document change permissions or trigger a write?

- [ ] **04 Check before execution**

  BUILD Authorize the concrete action at the service boundary, not just the prompt.

  TEST Does the service reject a substituted target or argument?

- [ ] **05 Keep delegated power bounded**

  BUILD Limit downstream authority to the original grant and current policy.

  TEST Can a child agent expand scope or revive a revoked grant?

- [ ] **06 Protect credentials and destinations**

  BUILD Use scoped credentials and constrain outbound connections.

  TEST Are wrong-audience tokens and unexpected destinations rejected?

Record evidence, an owner, and an unresolved issue for each check.

AI Agent Security Checklist

12 checks for design reviews and pre-production testing.

OPERATIONS & RECOVERY

- [ ] **07 Control durable memory**

  BUILD Record provenance and govern memory writes, access, and expiry.

  TEST Can untrusted content plant a persistent instruction?

- [ ] **08 Isolate execution**

  BUILD Bound filesystem, network, runtime, and resource consumption.

  TEST Can an allowed tool escape its assigned environment?

- [ ] **09 Bind consequential approvals**

  BUILD Require independent approval of the actual action and final parameters.

  TEST Does changing a recipient or amount invalidate the approval?

- [ ] **10 Capture useful evidence**

  BUILD Log action IDs, authorization outcomes, and relevant versions; redact secrets.

  TEST Can responders reconstruct the event without exposing credentials?

- [ ] **11 Test abnormal paths**

  BUILD Exercise hostile inputs, stale authority, tool errors, and repeated retries.

  TEST Do alerts and controls still work when the happy path breaks?

- [ ] **12 Rehearse containment and restart**

  BUILD Name who can stop execution, revoke access, and approve recovery.

  TEST Does a drill stop queued work and require a clean restart?

Practice guidance, not a certification. Tailor tests to the system and its risk.