AI Agent Sandbox Threat Model Checklist

AI Agent Sandbox Threat Model Checklist

Brand: OperatorForge AI
Updated: 2026-06-11
Format: free checklist + threat-model worksheet + consultant packaging notes
Use when: a founder, engineering lead, automation consultant, or ops team is deciding whether A

Why this matters now

The 2026-06-11 scan found a fresh cluster around AI agent sandboxing and accountable agent netw
- Hacker News surfaced "Your AI Agent Is Compromised" and "Principles of an Accountable AI Agen
- HN also surfaced a live discussion asking whether teams are defaulting to VM-level sandboxing
- GitHub search showed new or updated projects around semantic agent gateways, MCP threat scann
The monetizable gap: many teams are rushing straight to heavyweight VMs, browser isolation, or

15-minute agent sandbox threat model

Sandbox tier selector

Checklist

1. Inventory the agent’s blast radius

- Name each agent/tool and its owner.
- List every filesystem path, repository, SaaS workspace, and data class it can read.
- List every action that can change state: write file, run command, send message, update CRM,
- Mark payment, customer data, credentials, production infrastructure, and external communica

2. Choose the lightest safe isolation tier

- Do not start with "VM for everything"; start with the threat model and choose the minimum s
- Use Tier 0 only for non-sensitive, non-actioning work.
- Use Tier 1/2 when trusted agents need productivity but must respect path, command, network,
- Use Tier 3 for untrusted repositories, unknown MCP servers, downloaded scripts, or experime
- Use Tier 4 for irreversible or customer-visible actions.

3. Set read/write/tool boundaries

- Separate read permissions from write permissions.
- Keep secrets outside agent-readable paths and environment dumps.
- Require an explicit allowlist for shell commands, MCP tools, API endpoints, browser domains
- Deny credential discovery, token printing, bulk export, and hidden background daemons unles

4. Add approval gates before high-impact actions

- Require human approval before deploy, delete, refund, charge, email blast, data export, per
- Record the proposed action, risk tier, evidence, expected rollback, and approving owner.
- Use single-use approvals where possible; do not let one approval authorize a broad class of
- Make the kill switch accessible outside the agent chat/session.

5. Prove the controls work

- Run a harmless prompt-injection egress test.
- Try a blocked path, blocked command, and blocked network destination.
- Confirm logs capture agent ID, run ID, tool call, input summary, output summary, approval r
- Run one rollback drill before letting the agent touch production.

6. Package the result for owners

- Produce a one-page sandbox decision record.
- Include the selected tier, why heavier/lighter tiers were rejected, open risks, owner appro
- Keep a weekly list of blocked actions; blocked actions reveal either useful guardrails or f

Sandbox decision record template

Agent/tool:
Owner:
Primary job:
Data classes accessible:
State-changing actions allowed:
Selected sandbox tier:
Why this tier:
Explicit denials: