ZENTHOS

PRE-LAUNCH CHECKLIST

Agent Security Go / No-Go: 12 checks before an AI agent runs unattended

The rule (Meta, Oct 2025): an agent session may have at most two of untrusted input, sensitive data, and the ability to act externally. All three = no autonomy. Then check the twelve items below.

- [ ] **Trifecta statement written down**
  Which of the three properties this agent has in one session, and which one was removed, and how.
- [ ] **Agent has its own identity**
  A scoped service identity with short-lived tokens. Never a human's OAuth token or session.
- [ ] **Tool list is enumerated and least-privilege**
  Every tool named, read-only where possible, scoped to the task rather than the user.
- [ ] **Irreversible actions gated by a human**
  Money, deletion, publishing, permission changes need approval. Everything else does not.
- [ ] **Untrusted-content inventory**
  Emails, tickets, web pages, tool results: each source listed, with what is stripped before use.
- [ ] **Fetched content never picks tools**
  Two lanes: the user's instruction is privileged; anything retrieved mid-task is not.
- [ ] **Egress allowlist, reviewed**
  Domains the agent may reach are listed and checked for expired entries an attacker could buy.
- [ ] **Sandboxed, with dev/prod data separated**
  The agent cannot reach production data or credentials from a development or test session.
- [ ] **MCP servers and tools pinned and reviewed**
  Versions pinned, sourced from an internal registry, diffed on every update.
- [ ] **Action-level logging retained**
  Who, what, why, which tool, which data, for every run. Enough to reconstruct an incident.
- [ ] **Kill switch tested, owner named**
  One person can revoke the agent's credentials in minutes, and has practised doing it.
- [ ] **Blast-radius statement and red-team date**
  A written worst case for a successful prompt injection, and the next red-team scheduled.

12 of 12: run it. 9-11: fix the gaps first. Under 9: it is a pilot with a human in the loop, not a launch.