AI Agent PR CI Firewall Checklist

AI Agent PR CI Firewall Checklist

Brand: OperatorForge AI
Updated: 2026-06-14
Format: free checklist + CI policy template + evidence packet
Use when: a founder, engineering lead, agency, or AI-automation consultant lets Claude Code, Co

Why this matters now

The 2026-06-14 scan found a fresh cluster around AI-generated PRs needing deterministic gates i
- Hacker News surfaced "Show HN: Agent Gate - a deterministic CI firewall for AI-generated PRs"
- HN also surfaced "Superpowers for Claude, Codex etc.", "Hi HN: Loopy agent", "Double your Cod
- GitHub search found fresh repositories for agent security monitoring, MCP threat scanners, CI
- Reddit JSON checks for r/n8n, r/ClaudeAI, and r/LocalLLaMA returned HTTP 403 in this cron pro

The monetizable gap: teams are adopting autonomous coding agents faster than they are adding a

The 15-minute PR firewall rule

Every AI-authored pull request must pass a deterministic gate that is independent of the agent

AI Agent PR CI Firewall Checklist

1. Label agent-authored work

- Require agent PRs to include an ai-authored or agent-assisted label.
- Require a run ID, agent/tool name, model/provider if known, and human owner.
- Require a short scope statement: what the agent was allowed to change.
- Reject PRs with vague summaries such as "misc fixes" or "cleanup" when generated by an agen

2. Gate sensitive paths

- Require human approval if the diff touches auth, authorization, billing, payments, refunds,
- Block merge when a sensitive path is touched without a named reviewer.
- Require a rollback note for any production-risk file.
- Require a before/after diff summary tied to the original issue or requirement.

3. Prove tests were not skipped or weakened

- Capture targeted failing evidence before the fix when possible.
- Capture targeted passing evidence after the fix.
- Run the project baseline test/lint/typecheck command.
- Fail the PR if tests were deleted, weakened, renamed away from CI, or marked as skipped wit

4. Scan dependencies and commands

- Review package-lock, pnpm-lock, uv.lock, requirements, Dockerfile, shell scripts, GitHub Ac
- Require rationale for each new dependency.
- Block curl-to-shell, postinstall scripts, broad permission grants, unknown binary downloads
- Run a secret scan before merge.

5. Enforce evidence packet before merge

- PR includes run ID and owner.
- PR includes allowed scope and actual touched files.
- PR includes test commands and outputs.
- PR includes sensitive-path result.
- PR includes dependency diff result.
- PR includes rollback note or "not production-risk" rationale.
- Human reviewer checks the evidence; the agent cannot self-approve.

Minimal GitHub Actions policy sketch

```
name: ai-pr-firewall
on:
  pullrequest:
    types: opened, synchronize, reopened, labeled, unlabeled
jobs:
  firewall:
    runs-on: ubuntu-latest
    steps:
      - uses: actions/checkout@v4
```