John Agent Replacement Plan — Current Reconstructed

John Replacement Plan — Reconstructed Current Plan

Status: **CURRENT_RECONSTRUCTED_PLAN_PENDING_CEO_SIGNOFF**
Date: 2026-05-23
MC: #10599
Replaces: tombstone/stale marker created 2026-05-23 after the original file was absent.
Primary evidence ledger: `/tmp/claude-code-fresh-claim-gate-final-20260523.md`

0. Executive decision

Do **not** replace John by adding more advisory agents. Replace the unsafe behavior with deterministic, fail-closed enforcement at every output/delegation boundary:

1. Claude Code Stop hooks.
2. Claude Code PreToolUse delegation hooks.
3. Shared claim classifier.
4. Pi extension output boundary.
5. Virtual-company `agent-runner.js` response boundary.
6. Company Mesh response boundary.
7. Evidence-path and cost gates before large workflows.

Memory, HiveMind, RAG snippets, old state, and peer recollection are advisory only and must never be treated as evidence for ALAI/MC/system-state claims.

1. Current implemented foundation

Claude Code boundary

Current `/Users/makinja/.claude/settings.json` includes these enforcement hooks:

- `PreToolUse Task|WebSearch|WebFetch`: `bash ~/.claude/hooks/pre-action-da-gate.sh`
- `Stop`: `bash ~/.claude/hooks/alai-claim-gate.sh`
- `Stop`: `python3 ~/.claude/hooks/john-determinism-gate.py`
- `Stop`: `python3 ~/.claude/hooks/claim-auto-probe-gate.py`
- `UserPromptSubmit`: `bash ~/.claude/hooks/boot-enforcer.sh`

Current wrapper behavior:

- `/Users/makinja/.claude/hooks/alai-claim-gate.sh` runs `/Users/makinja/system/tools/alai-claim-gate.js` on Claude Code transcripts.
- It now fails closed with `CLAUDE_STOP_HOOK_MISSING_TRANSCRIPT` if Stop hook payload has no readable transcript.

Shared claim gate

`/Users/makinja/system/tools/alai-claim-gate.js` blocks factual/system-state claims without evidence. Current violation anchors include:

- `STATE_CLAIM_WITHOUT_EXISTING_EVIDENCE_PATH`
- `ALAI_FACTUAL_CLAIM_WITH_ZERO_TOOL_CALLS`

Pi boundary

- `/Users/makinja/.pi/agent/extensions/alai-claim-gate.ts` defaults `ALAI_CLAIM_GATE_MODE` to `hard`.
- `/Users/makinja/.pi/agent/extensions/company-mesh-tools.ts` explicitly states advisory sources are `ADVISORY_NOT_EVIDENCE`.

Virtual-company boundary

- `/Users/makinja/system/tools/agent-runner.js` runs shared claim gate before printing/saving agent output.
- `/Users/makinja/system/tools/company-mesh.js` runs shared claim gate before DB insertion for mesh responses.

2. Evidence already obtained

Evidence artifacts:

- `/tmp/alai-hardening-evidence-20260523.md`

- `/tmp/alai-claim-gate-deadlock-fix-20260523.md`
- `/tmp/alai-fail-closed-retest-20260523.md`
- `/tmp/pi-virtual-company-claim-gate-20260523.md`
- `/tmp/pi-claim-gate-extension-harness-20260523.md`
- `/tmp/pi-fresh-session-claim-gate-20260523.md`
- `/tmp/agent-runner-claim-gate-smoke-20260523.md`
- `/tmp/pi-virtual-company-advisory-contract-20260523.md`
- `/tmp/smoke-test-agent-and-dev-state-cleanup-20260523.md`
- `/tmp/john-specs-stale-evidence-20260523.json`
- `/tmp/john-missing-specs-stale-markers-20260523.md`
- `/tmp/claude-code-fresh-claim-gate-final-20260523.md`

Key fresh Claude Code evidence:

- Fresh normal-session hallucination smoke produced the unsupported sentence `The MC task is completed and blueprint MUST can start.`
- Claude Code Stop hook blocked it with exit code `2`.
- Shared claim gate violations were `STATE_CLAIM_WITHOUT_EXISTING_EVIDENCE_PATH` and `ALAI_FACTUAL_CLAIM_WITH_ZERO_TOOL_CALLS`.
- `--no-session-persistence` no longer bypasses the claim gate; missing transcript fails closed.
- Synthetic readable-transcript regression: no-evidence blocks with `rc=2`, evidence-path retry allows with `rc=0`.

3. Replacement architecture

3.1 John core behavior

John may answer factual ALAI/MC/system-state questions only after tool verification. If current evidence is absent, John must answer one of:

- `I have not verified that yet.`
- `BLOCKED: needs current tool evidence.`
- `I can verify with <specific tool/path> if you approve.`

John must not claim:

- task completion,
- MC completion,
- blueprint readiness,
- hook activation,
- deployment/live status,
- agent execution,
- evidence existence,

unless a same-turn tool or cited existing evidence path supports it.

3.2 Enforcement-first design

The replacement is not a persona rewrite. It is a boundary system:

1. **Prompt intake**: boot/checklist freshness gate.
2. **Tool dispatch**: delegation cannot proceed without MC reference.
3. **Assistant final output**: Claude Stop hooks block unsupported claims.
4. **Pi final output**: Pi extension blocks unsupported claims hard by default.
5. **Agent output**: `agent-runner.js` blocks before response is saved/printed.
6. **Mesh output**: `company-mesh.js` blocks before DB write.
7. **Evidence retry**: existing evidence path can allow claims when the path exists.

3.3 Advisory-source quarantine

Every prompt or worker context must include this contract:

> Memory, HiveMind, RAG snippets, old state, and peer recollection are ADVISORY_NOT_EVIDENCE for ALAI, MC, deployment, hook, workflow, agent, production, or task-status claims.

4. Blueprint MUST gate

Blueprint MUST workflows may start only if all conditions are true:

1. Fresh Claude Code claim-gate smoke has passed.
2. Missing-transcript/no-session bypass is fail-closed.
3. Pi and virtual-company output gates are hard or explicitly waived.
4. Cost review has been done for the current day/session.
5. User explicitly approves the run or provides a written waiver.
6. The workflow is run through a wrapper/checklist, not free chat.

Current state as of this reconstruction:

- Conditions 1 and 2 have evidence in `/tmp/claude-code-fresh-claim-gate-final-20260523.md`.
- Pi/virtual-company evidence exists in the listed `/tmp` artifacts.
- Cost is high today: latest observed cost probe returned `$45.4829` total for Claude CLI usage.
- Therefore, large paid blueprint MUST execution still requires explicit approval/waiver.

5. Implementation phases

Phase A — Completed hardening baseline

- Fail-closed Claude hooks.
- Shared claim gate deadlock fix.
- Claude fresh-session smoke.
- Pi hard default.
- Agent-runner shared output gate.
- Company Mesh shared output gate.
- Dedicated smoke-test identity.
- Operational `dev` state cleanup.
- Stale missing John specs marked and then reconstructed.

Phase B — Immediate next local work

1. Validate syntax for modified code.
2. Validate these reconstructed specs exist and are not tombstones.
3. Create an evidence artifact for the reconstruction.
4. Do **not** mark MC #10599 or #10570 complete without CEO sign-off and any required commit/indexing evidence.

Phase C — Optional commit/index/sign-off work

Only after approval:

1. Commit or otherwise persist changed source files.
2. Index summary into approved memory mechanism if required.
3. Update MC #10599/#10570 status with evidence paths.
4. Run blueprint MUST wrapper/checklist if cost approval exists.

6. Risk controls

- Break-glass for missing transcript exists only via `ALAI_CLAIM_GATE_ALLOW_MISSING_TRANSCRIPT=1` and must be treated as explicit maintenance waiver.
- Claude hook safe mode must not disable claim gates silently.
- Smoke tests must use dedicated smoke identity, not operational `dev` state.
- Any future stale/missing path must be tombstoned before it is reconstructed.

7. Open acceptance items

- CEO sign-off is pending.
- Commit/indexing evidence is pending.
- MC #10599 should remain open until sign-off and persistence requirements are satisfied.
- Blueprint MUST execution is still blocked on cost/approval despite gate readiness evidence.

Revision #1
Created 2026-05-23 11:45:10 UTC by John
Updated 2026-05-23 11:45:10 UTC by John