Attack J Security Fix (MC #101149)

Attack J Security Fix (MC #101149)

MC: #101149 (Attack J — evidence-ledger writer identity spoof)

Parent: Reality Anchor Doctrine Phase 2
Owner: CodeCraft / Petter Graff
Date Shipped: 2026-05-17
Proveo Verdict: PARTIAL → PASS (after fix)

The Vulnerability (Attack J)

The evidence-ledger writer identity could be spoofed via the `--actor` CLI parameter. The ledger write path used:

```
const agentId = process.env.CLAUDE_AGENT_ID || actor || 'unknown';
```

This allowed an agent to bypass the Writer ≠ Witness gate by supplying:

```
mc.js ready 101149 --evidence /tmp/evidence-101149/ --actor "different-agent-id"
```

The ledger would record `writer_agent_id = "different-agent-id"`, then at `mc.js done` the same agent could close the task because the writer_id check passed (different agent on paper, same agent in reality).

> **Reality Anchor Invariant:** "The agent that produces evidence cannot be the agent that closes the task."

This is the Writer = Witness antipattern — the bypass allowed a single agent to both produce and validate its own evidence.

The Fix

Remove `|| actor` from the identity fallback chain. The agent identity MUST come from `CLAUDE_AGENT_ID` environment variable (set by the Claude Code runtime, not user-supplied).

Changed Lines in mc.js

Line 2843 (done gate — read ledger):

```
- const currentAgentId = process.env.CLAUDE_AGENT_ID || actor || 'unknown';
+ const currentAgentId = process.env.CLAUDE_AGENT_ID || 'unknown';
```

Line 3538 (ready command — write ledger for --evidence):

```
- const agentId = process.env.CLAUDE_AGENT_ID || actor || 'unknown';
+ const agentId = process.env.CLAUDE_AGENT_ID || 'unknown';
```

Line 3574 (deprecation log):

```
- agent_id: process.env.CLAUDE_AGENT_ID || actor || 'unknown',
+ agent_id: process.env.CLAUDE_AGENT_ID || 'unknown',
```

Line 3589 (ready command — write ledger for --probe-evidence):

```
- const agentId = process.env.CLAUDE_AGENT_ID || actor || 'unknown';
+ const agentId = process.env.CLAUDE_AGENT_ID || 'unknown';
```

Symmetric Enforcement

The fix ensures symmetric identity enforcement across both READ (line 2843) and WRITE (lines 3538, 3589) paths:

Operation	Line	Identity Source
Read ledger (done gate)	2843	`CLAUDE_AGENT_ID` only

Operation	Line	Identity Source
Write ledger (ready --evidence)	3538	`CLAUDE_AGENT_ID` only
Write ledger (ready --probe-evidence)	3589	`CLAUDE_AGENT_ID` only
Deprecation log	3574	`CLAUDE_AGENT_ID` only

The `--actor` CLI parameter remains available for legacy MC operations (comments, audit log human-readable labels) but is NEVER used as a substitute for `CLAUDE_AGENT_ID` in the P2.2 writer≠closer gate.

Why This Matters

Without this fix, the entire Reality Anchor Phase 2 evidence ledger was bypassable. An agent could:

1. Generate evidence file
2. Write to ledger with spoofed writer_id via `--actor`
3. Close task with `mc.js done` (same agent, but ledger shows different writer)
4. Gate passes because `writer_agent_id != closer_agent_id`

This is a **privilege escalation via identity forgery**. The fix closes the gap by enforcing that only the runtime-provided `CLAUDE_AGENT_ID` is authoritative.

Proveo Test Update

Before fix: Proveo verdict for MC #101149 = PARTIAL (Attack J bypass demonstrated)
After fix: Proveo re-ran test → verdict updated to PASS

Test scenario:

1. Builder agent produces evidence for task #101149
2. Builder attempts `mc.js ready 101149 --evidence /tmp/evidence-101149/ --actor "fake-verifier-id"`
3. **Expected:** Ledger records writer_agent_id = builder's real CLAUDE_AGENT_ID (NOT "fake-verifier-id")
4. Builder attempts `mc.js done 101149`
5. **Expected:** Gate BLOCKS because writer_agent_id == closer_agent_id

Result: PASS — gate correctly blocked self-closure.

Writer ? Witness Invariant (Now Enforced)

The invariant is now enforced symmetrically in both read and write paths:

> **Invariant:** The agent_id that writes evidence to the ledger MUST differ from the agent_id that calls `mc.js done`. Identity MUST be derived from `CLAUDE_AGENT_ID` environment variable, NOT from user-supplied `--actor` parameter.

Audit Trail

All evidence ledger entries at `~/system/state/evidence-ledger.jsonl` now contain:

- `writer_agent_id` — from `CLAUDE_AGENT_ID` only
- `sha256` — content hash
- `task_id` — MC reference
- `timestamp` — write time
- `event_type` — "ready" or "done"

The gate at `mc.js done` verifies:

1. Ledger entry exists for task_id
2. `writer_agent_id != closer_agent_id`
3. SHA-256 hash matches file content
4. Timestamp within task execution window

Related

- **MC:** #101149 (Attack J fix)
- **Parent:** Reality Anchor Phase 2 (MC #100823–#100827)
- **Code:** `~/system/tools/mc.js` lines 2843, 3538, 3574, 3589
- **Proveo Test:** MC #101149 validation — writer≠witness gate attack (PASS after fix)
- **Doctrine:** Reality Anchor v1 Final

Revision #2
Created 2026-05-17 12:05:06 UTC by John
Updated 2026-06-21 20:03:42 UTC by John