Bilko Sentinel — Tier-1 Bounded Auto-Remediation (Shadow-First) 2026-06-11

Status

BUILT — SHADOW mode. MC #103435 (AgentForge build) + MC #103436 (Securion adversarial review). Parent MC #103328. Module: `/Users/makinja/system/tools/bilko-sentinel-tier1.js`. **Tier-0 remains the active detection layer; Tier-1 is shadow-calibrating only.**

Related: Bilko Sentinel Tier-0 (detection layer, LIVE) — Bilko Observability (GCP-native)

SRE rationale: DECISIONS-observability-2026-06-10.md, Decision 2 — Kelsey Hightower consult. The bar is not yet met; ship the muscle, start the calibration clock, arm only when proven.

What Tier-1 Is

Tier-1 is the agent that can **fix**. Where Tier-0 only detects, diagnoses, and proposes — Tier-1 computes the exact bounded action and, when armed, executes it. Permitted action set (exhaustive):

- Roll back `bilko-api-demo` or `bilko-web-demo` to N-1 only (never older)
- Set Cloud Run `--min-instances` 0→1 (warm floor for cold-start incidents)
- Slack escalation (always permitted, labelled with current mode)

Never-automate (enforced at IAM, not policy): no IAM/policy change, no Cloud SQL op, no secret op, no DNS/LB/network, no rollback older than N-1, no action during an in-flight deploy or protected business window. Invoked by the Tier-0 loop via try/catch — any Tier-1 error is caught and logged, Tier-0 continues unaffected.

Modes (SENTINEL_TIER1_MODE)

Mode is read **once at startup**, immutable for the life of the process. Unrecognised or missing value resolves to `shadow`. The LaunchAgent plist deliberately omits the key.

Mode	Behaviour	Current state
shadow (DEFAULT)	On a breach: compute the bounded action, announce to #ceo with gate evaluation, write one row to calibration ledger. Execute nothing.	ACTIVE
`ack`	Same as shadow, but if a human posts APPROVE in the #ceo thread within 3 min, execute the action (all 8 gates must pass). Slack poll loop is a follow-on — currently defers conservatively. Requires F4 hardening first.	NOT YET WIRED
`auto`	Execute automatically when all 8 gates pass and promotionBarMet() returns true. Silence in ack window = proceed. FORBIDDEN until bar met + human-engineer sign-off.	BLOCKED (PROMOTION_BAR_NOT_MET)

Current State: SHADOW — Confirmed Inert

Securion adversarial review (MC #103436) + AgentForge build verification (MC #103435) independently confirm: **in shadow mode this agent cannot mutate prod.** Two independent structural barriers:

1. `handleIncident()` enters an `if (MODE === 'shadow')` block and **returns at line 852** — the execution block at line 861+ is outside and structurally unreachable.
2. `executeRollback()` (line 625) and `executeScaleFloor()` (line 675) each **throw as their literal first statement** when MODE === 'shadow'. Both barriers are independently sufficient.

Live production traffic verified unchanged during shadow simulation: `bilko-api-demo-00192-sfv @ 100%` before and after. Auto mode tested with `SENTINEL_TIER1_MODE=auto` and hard-blocked with PROMOTION_BAR_NOT_MET.

8 Pre-fire Gates

ALL eight must be true before any execution in ack or auto mode. In shadow they are evaluated and recorded to the ledger only.

#	Gate	Detail

1	Alert sustained ≥5 min	Prevents action on transient spikes. Measured from `incident.firstSeenAt`.
2	Calibrated LLM confidence	Requires "high" until ≥5 ledger reviews; adjusts to "medium" with ledger data. Derived from calibration, not hardcoded.
3	N-1 confirmed healthy ≥10 min	Rollback target must have been in Ready=True state for ≥10 min before the current (bad) revision. Unknown = block + escalate.
4	No schema migration in bad revision	Requires deploy manifest (`~/system/state/bilko-deploy-manifest.json`). **Honest block: absent manifest = BLOCK + escalate to human.** Rolling back across a schema migration can corrupt data.
5	Cooldown: no action in last 60 min	One action per 60-minute window across all types.
6	3-min human-ack window	HOLD or ABORT in #ceo thread cancels. Shadow: informational. Ack: requires explicit APPROVE. Auto: silence = proceed.
7	IAM diff vs known-good snapshot	Compares live Cloud Run service IAM policy against `~/system/state/bilko-sentinel-iam-snapshot.json`. Mismatch = block + escalate. Motivated by the 2026-06 IAM wipe incident.
8	N-1 is not itself a rollback revision	Prevents rolling back to a revision tagged as a known-bad rollback. Checked against deploy manifest `isRollback` flag.

Circuit Breakers

- **Max 2 actions / 24h** across all types. Third incident in 24h → human-only escalation.
- **Self-disable after failed remediation:** post-action health check at +5 min; if service still unhealthy → `circuitOpen=true`, SENTINEL-CIRCUIT-OPEN to Slack + email. Manual re-enable: set `circuitOpen=false` in `~/system/state/bilko-sentinel-tier1-state.json`.
- **Single-writer lock:** atomic file lock (`~/system/state/bilko-sentinel-tier1.lock`) — prevents race on same revision.
- **Audit-before-execute:** `~/system/logs/bilko-sentinel-audit.jsonl` written before any gcloud mutation verb. Log write failure → action does not fire.
- **Two Slack announcements per action:** BEFORE ("about to roll back X to rev Y in 3 min unless HOLD") and AFTER ("rolled back, health check in 5 min").

Promotion Bar: shadow ? auto

`promotionBarMet()` is a hard gate on the auto path. Reads the calibration ledger at runtime — not hardcoded. Currently returns **FALSE**; auto path refuses with PROMOTION_BAR_NOT_MET.

Criterion	Required	Current (2026-06-11)
daysLive_gte30	≥30 days since first ledger entry	0 days — NOT MET
evaluatedProposals_gte20	≥20 proposals in ledger	2 — NOT MET
fpRate_lt5pct	Human-reviewed FP rate <5%	100% default — NOT MET
groundTruthHit	≥1 row with human_verdict=correct	0 — NOT MET
deployManifestExists	~/system/state/bilko-deploy-manifest.json present	Absent — NOT MET

When the bar is eventually met, **human-engineer sign-off is still required** before the plist is updated to set `SENTINEL_TIER1_MODE=auto`. That is an explicit audited step, not an automatic promotion.

Arming Prerequisites (Securion #103436 — MC #103439)

These gate **arming (ack/auto), not shadow.** Securion re-review required before the mode key is added to the plist.

Finding	Severity	Required before arming
F5 — Ledger integrity	HIGH	HMAC-sign each ledger row (key stored outside ledger path). promotionBarMet() must verify HMAC before counting. Without this, forged human_verdict entries can satisfy the promotion bar and arm auto.
F7 — SA IAM scope	MEDIUM	Verify alai-cli-deployer holds only monitoring.viewer + logging.viewer + run.viewer in shadow. For auto: run.developer scoped by resource condition to bilko-api-demo + bilko-web-demo only. Must NOT hold cloudsql.*, iam.*, secretmanager.*, dns.*.
F4 — Ack approver allowlist	INFO	Before Slack poll loop is wired: define constant with allowed Slack user IDs. Poll must verify message.user + thread_ts — not just message text.

Finding	Severity	Required before arming
F6 — IAM snapshot seal	MEDIUM	chmod 0444 after first write. Add sealed flag; require manual unsealing for reset. Populate bilko-web-demo immediately.
F2 — Misleading Object.freeze	MEDIUM	Remove Object.freeze({MODE}) at line 57 — it freezes a discarded object, not the const binding. Replace with clarifying comment.
F8 — Gate 8 inconsistency	LOW	Align Gate 8 with Gate 4: block (not warn-and-pass) when deploy manifest absent or N-1 not in manifest.
F3 — Module integrity	LOW	Add startup SHA-256 check of the module file against a stored known-good value outside the module path.

All items tracked in MC #103439. **Securion re-review required before mode key is added to plist.**

Calibration Ledger

Every shadow proposal appends one row to `/Users/makinja/system/logs/bilko-sentinel-tier1-ledger.jsonl`

Row schema: `{ ts, incidentId, policyName, condName, resource, diagnosis, confidence, computedAction, n1Info, gates, mode, human_verdict }`

The `human_verdict` field starts as `"not-yet-reviewed"`. Update to: `correct` | `wrong-rootcause` | `would-have-worsened`. A **weekly summary** is posted to #ceo automatically: proposal count, reviewed count, FP rate, ground-truth hits, promotion bar status.

```
node /Users/makinja/system/tools/bilko-sentinel-tier1.js --weekly-summary
```

Infrastructure

Component	Location
Module	`/Users/makinja/system/tools/bilko-sentinel-tier1.js`
Weekly summary plist	`/Users/makinja/system/tools/com.alai.bilko-sentinel-tier1-weekly-summary.plist`
Calibration ledger	`/Users/makinja/system/logs/bilko-sentinel-tier1-ledger.jsonl`

Component	Location
IAM snapshot	`/Users/makinja/system/state/bilko-sentinel-iam-snapshot.json`
State file	`/Users/makinja/system/state/bilko-sentinel-tier1-state.json`
Execution audit log	`/Users/makinja/system/logs/bilko-sentinel-audit.jsonl`
Run log	`/Users/makinja/system/logs/bilko-sentinel-tier1.log`
Single-writer lock	`/Users/makinja/system/state/bilko-sentinel-tier1.lock`
GCP project	`tribal-sign-487920-k0`, region `europe-north1`
SA	`alai-cli-deployer@tribal-sign-487920-k0.iam.gserviceaccount.com`
Allowed services	`bilko-api-demo`, `bilko-web-demo`
Host	ANVIL (makinja local Mac) — invoked by Tier-0 loop

Runbook

Self-test in shadow

```
node /Users/makinja/system/tools/bilko-sentinel-tier1.js --self-test
```

Evaluate promotion bar

```
node /Users/makinja/system/tools/bilko-sentinel-tier1.js --promotionbar-test
# exits 0 if bar met, 1 if not
```

Inspect calibration ledger

```
tail -f /Users/makinja/system/logs/bilko-sentinel-tier1-ledger.jsonl
```

Re-enable after circuit-open

```
# Edit ~/system/state/bilko-sentinel-tier1-state.json
# Set "circuitOpen": false
```

Flip to ack mode (AFTER all hardening items + Securion re-review)

```
# Add to LaunchAgent plist EnvironmentVariables:
#   <key>SENTINEL_TIER1_MODE</key><string>ack</string>
launchctl unload ~/Library/LaunchAgents/com.alai.bilko-sentinel.plist
launchctl load ~/Library/LaunchAgents/com.alai.bilko-sentinel.plist
```

Revision #1
Created 2026-06-11 11:57:33 UTC by John
Updated 2026-06-11 11:57:33 UTC by John