



JUDGEMENT SPINE
— Human. AI. Judgement. Covered —

Security Model - Threat Model Summary (Public)

A minimal threat model for execution-time governance systems.

Threats addressed

- Authority drift (agents expanding capability mid-run).
- Bypass of approvals via alternate paths (shadow execution).
- Evidence tampering after incidents (log manipulation).
- False positives at scale (automated containment / batch commits).

Controls

- Inline interception at commit boundaries.
- Explicit authority chain + expiring approvals.
- Bounds written into execution mechanisms.
- Tamper-evident evidence packs (hash manifest + signatures).

Residual risks

- If a boundary cannot be intercepted, it cannot be governed (must redesign control surface).
- If approver rosters are wrong, escalation routes will fail (keep rosters current).
- If downstream systems ignore bounds, enforcement must move closer to commit.

Contact: founder@judgementspine.com