Risk, Trust, and Safety Model

How we keep signals trustworthy, keep operators safe, and keep alerts human■centered.

1) Source Provenance & Integrity

• Maintain source allowlist (CVE/NVD, CISA/CERT, vendor/OEM) and denylist for known spoofers.
• Record provenance on every alert: origin URL, timestamp, TLP, hash, and a confidence score.
• Detect poisoning/spoofing: domain verification, signature/PGP where available, verbatim match checks.
• Unverified tips enter a queue for Tier■2 analyst validation; never auto■escalate.

2) User Verification (Lightweight)

• Invitation or referral-based onboarding; organization email and role attestation.
• Optional backchannel thumbs-up from trusted partners (where available).
• Risk-based access: sensitive guidance visible only to verified operators.

3) AI Safety & Human Oversight

• AI summarizes long advisories and proposes a score; humans confirm/adjust before action.
• Guardrails: instruction tuning against “unsafe actions” (no one■click shutdowns); prompt■injection tests.
• Log AI outputs and human overrides for error analysis; maintain a denylist of unsafe recommendation phrases.

4) Alert Fatigue & Notification Hygiene

• Mission■control single■card view; limit concurrent active alerts to the top few.
• Quiet Mode: only human■safety (RED) alerts can break through during protected windows.
• Suppression: duplicates, non■applicable alerts (via “Is this us?”), and low■confidence sources.

5) Auditability & Governance

• Every GO/HOLD/ESCALATE decision is timestamped with user ID, score breakdown, and provenance receipt.
• Tier■2 escalation packages the alert, environment context, and logs for follow■up.
• Periodic review: sample AI summaries, false positives/negatives, and notification break■throughs.

Outcome: trustworthy signals, fewer alerts, safer next steps — aligned with Bio■ISAC’s mission and constraints.