The 3 A.M. Rule

What your AI security tools may and may not do without you

Your AI tools may watch, sort, and draft. A person approves anything that disconnects, deletes, or locks people out.

WHY THIS EXISTS

Security vendors now sell agents that respond to attacks automatically. The attacks they face no longer wait for a human: one style of intrusion probes quietly for weeks to learn your habits, then strikes during a shift change[1], and the instructions that trigger it can arrive inside an ordinary document or support ticket[2].

The real failure mode is not that the tool gets hacked. It is that the tool acts on a guess. When an agent misreads a routine overnight backup as a theft in progress, it can cut off the network itself and cause the outage with no attacker involved[3].

THE 60-SECOND GATE

Before you approve any containment action, answer four questions.

1. **Second signal?** Two independent tools agreeing, or a person reporting it? One alert alone means wait.
2. **Critical host?** If the machine runs billing, scheduling, or patient records, a second person decides with you.
3. **Where did it start?** An opened attachment, a clicked link, pasted text into a chatbot or ticketing tool means treat it as an injection, not just malware.
4. **What breaks if we are wrong?** If a wrong call costs a customer a day of work, escalate instead of approving.

THREE NON-NEGOTIABLES

1. **May.** Read logs, rank alerts, draft a containment plan, open the change ticket, notify staff.
2. **May not.** Disable a network port or VPN, revoke accounts, delete or rebuild a machine, or shut down a server without a named person confirming.
3. **Must.** Name that person in writing before the tool is switched on, and name a backup for nights, weekends, and vacations.

COPY THIS INTO YOUR POLICY

SECTION: AI AND AUTOMATED SECURITY TOOLS

No AI or automated security tool may take an action that interrupts production, including disabling network access, revoking credentials, deleting data, or shutting down a host, without confirmation from a named human operator. Automated tools may collect data, prioritize alerts, and draft recommended actions. At least one operator and one backup operator must be named in this policy. Every automated containment action must be logged with the alert that triggered it and the person who approved it.

IF IT WAS A FALSE ALARM

- Restore service first. Write the incident up second.
- Record what the tool saw, what was actually happening, and who approved.
- Raise the threshold that fired. If the same rule fires twice in one week, delete it rather than keep tuning it.

START HERE IF YOU HAVE NOTHING WRITTEN

Open a single document called *security policy*. Name one operator and one backup. Paste the clause above underneath those names. Save it where the on-call phone lives.

SET THESE TWO THRESHOLDS

Automatic actions are allowed only on non-production devices, and only during staffed hours. Everything else waits for a person, even at 3 a.m.

WHAT THIS CARD DOES NOT DO

It does not replace a lawyer, an insurer, or an incident response retainer. It does not promise compliance. It gives one non-negotiable rule, in writing, before the tool is switched on.

Sources. 1 Bougioukas, "How 'Silent Probing' Can Make Your Security Playbook a Liability," CyberScoop, 2026. **2** Das, Amini, and Wu, "Security and Privacy Challenges of Large Language Models," ACM Computing Surveys, 2025. **3** Sullivan and Callow, "Ghost Breaches," CyberScoop, 2026.

Print this. Pin it where the on-call phone lives. Full evidence base and citations: **english.scushi.dev**. Student project for ENGL 102. Not legal, compliance, or professional security advice.