MERCY REGISTRY / FIELD PROTOCOL MK-FM-01

AI INCIDENT FIELD MANUAL

A practical, vendor-neutral response kit for when an AI system, agent, or automation does something it should not.

CONTAIN	VERIFY	RECOVER
Stop further action	Establish facts	Return safely

FREE SAMPLE / 6 PAGES
FOR INDIE BUILDERS, OPERATORS, AND SMALL TEAMS

Use this manual for operational readiness. It is not legal, cybersecurity, medical, or emergency-services advice.

00 / ORIENTATION

What this manual is - and is not.

This is a fast operational aid for bounded AI incidents. It helps you reduce harm, preserve evidence, and make a deliberate recovery decision.

USE IT WHEN

An agent acts without approval Messages, purchases, deletes, deployments, or account changes occur outside the intended scope.	**Instructions are hijacked** External content, retrieved text, or a user prompt alters tool behavior or bypasses policy.
Sensitive data appears Secrets, personal data, internal files, or private conversation content may have been exposed.	**Bad output ships** Incorrect, harmful, or unreviewed AI output reaches customers, coworkers, or production systems.

OPERATING PRINCIPLE

Contain first. Preserve evidence second. Communicate facts third. Recover only after the unsafe path is understood and blocked.

BOUNDARIES

- ☐ If a person may be in immediate danger, contact local emergency services and follow your established emergency plan.
- ☐ If regulated data, fraud, extortion, or material financial loss may be involved, escalate to qualified legal, security, and financial professionals.
- ☐ Do not destroy logs, conceal impact, or promise an outcome before the facts are known.
- ☐ Do not let the same system investigate itself as the only source of truth.

Framework note: the structure follows the spirit of NIST AI RMF functions - govern, map, measure, and manage - and OWASP guidance on prompt injection and excessive agency. See Sources on the final page.

01 / FIRST 60 SECONDS

Stop the blast radius.

Do these in order. If one step would erase evidence or create more risk, pause and involve the incident owner.

STOP
Pause the agent, workflow, scheduled job, deployment, or integration. Use the fastest reversible control available.

CUT PRIVILEGE
Revoke or narrow tool access, tokens, API keys, write scopes, payment authority, and external messaging permissions.

PRESERVE
Save timestamps, prompts, model outputs, tool calls, request IDs, screenshots, logs, affected object IDs, and the current configuration.

BOUND
Identify what the system could read, change, send, buy, delete, or publish during the exposure window.

ASSIGN
Name one incident lead and one note-taker. Route decisions through the incident lead.

ESCALATE
If people, money, regulated data, or production integrity may be affected, use the highest relevant severity while facts are incomplete.

ONE SENTENCE TO SAY OUT LOUD

We have paused the affected automation, preserved the available evidence, and are determining scope before we restore service.

DO NOT

Do not rerun the prompt A rerun can repeat the harmful action or overwrite useful state.	**Do not rotate everything blindly** Unplanned rotation can lock out responders or erase attribution. Prioritize exposed credentials.
Do not delete the evidence Preserve originals. Work on copies and record every containment change.	**Do not speculate publicly** State what is known, what is being checked, and when the next update will arrive.

02 / SEVERITY

Classify by impact, not embarrassment.

Choose the highest row that might reasonably apply. Downgrade only when evidence rules it out.

LEVEL	SIGNALS	RESPONSE
SEV-1	Safety risk; active fraud; regulated data exposure; destructive production action; widespread unauthorized external action.	Stop affected systems. Activate executive, legal, security, and continuity response now. Update at a fixed cadence.
SEV-2	Material customer impact; limited sensitive data exposure; unauthorized write action; recurring uncontrolled behavior.	Assign incident lead. Contain within minutes. Notify accountable owners. Prepare customer communication.
SEV-3	Wrong or harmful output with bounded reach; reversible internal changes; no evidence of sensitive data exposure.	Pause workflow. Investigate same day. Correct affected outputs and add a targeted control before restart.
SEV-4	Near miss; blocked prompt injection; test-only anomaly; no unauthorized action or external impact.	Document, reproduce safely, fix, and add a regression test. Share the lesson with operators.

FAST SCOPE QUESTIONS

- [ ] What is the earliest confirmed bad action? What is the latest possible bad action?
- [ ] Which identities, tools, models, data stores, destinations, and payment methods were reachable?
- [ ] Was the behavior caused by user input, retrieved content, model output, configuration, code, or credentials?
- [ ] What independent record confirms each claim: provider log, database history, audit trail, email receipt, or human witness?
- [ ] What would make the impact one level worse? Check that condition explicitly.

DEFAULT RULE

When facts are missing, classify provisionally at the higher plausible severity.

03 / PLAYBOOK A

Runaway agent or unauthorized action.

Use when an AI agent sends, purchases, edits, deletes, deploys, or invokes tools outside the intended scope.

CONTAIN

- [ ] Disable the specific workflow or agent before disabling unrelated services.
- [ ] Revoke its session and write-capable credentials. Preserve credential IDs and rotation timestamps.
- [ ] Block outbound messages, transactions, deployments, and destructive actions at the downstream service.
- [ ] Freeze scheduled retries, queues, and background workers that can replay the action.

VERIFY

- [ ] Export the complete tool-call chain, including arguments, responses, failures, and retries.
- [ ] Compare intended task scope with actual permissions and actual actions.
- [ ] Check downstream audit logs. The agent transcript alone is not proof of what occurred.
- [ ] Enumerate every object touched and whether the change is reversible.

RECOVER

- [ ] Restore from a known-good state or reverse each confirmed action with an auditable change.
- [ ] Reduce functionality, permissions, and autonomy to the minimum required.
- [ ] Require human confirmation for external messages, spending, deletion, permission changes, and production deploys.
- [ ] Add budgets, rate limits, idempotency keys, dry-run mode, and a tested kill switch.

EXIT CONDITION

Do not restart until the original unsafe path fails in a controlled regression test.

SAMPLE / CONTINUE

You now have the first-response core.

The full manual adds five incident playbooks, communication templates, evidence and vendor checklists, recovery gates, a postmortem template, a pre-flight review, and printable worksheets.

5 playbooks Runaway agent, prompt injection, credential exposure, sensitive data, and harmful output.	**2 response templates** Internal update and vendor escalation formats designed for fast, factual communication.
3 worksheets Timeline, incident command sheet, and a cut-and-fold response card.	**Preventive controls** A launch checklist covering authority, bounds, evidence, and testing.

FULL EDITION

Read all 18 pages free for evaluation. US$19 plus applicable tax buys a one-organization internal-use license. No subscription.

GET THE FULL MANUAL

Visit: https://mercy-key.codex507411.chatgpt.site/field-manual

This sample and the full manual are educational operational aids, not professional advice. See the full edition for complete sources, license, and disclaimer.