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

FULL EDITION / VERSION 1.1
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.

04 / PLAYBOOK B

Prompt injection or tool hijack.

Use when content from a document, website, email, message, database, or user causes an AI system to ignore intended instructions or misuse tools.

CONTAIN

- [ ] Disable the affected retrieval source, connector, browser action, or tool path.
- [ ] Quarantine the triggering content without opening or reprocessing it through the same automation.
- [ ] Revoke any credential the system exposed, echoed, or used outside expected destinations.

VERIFY

- [ ] Preserve the exact untrusted input, system instructions, model/version, tool schema, output, and tool trace.
- [ ] Determine whether the injected instruction was merely repeated or actually changed a tool action.
- [ ] Check whether the same content exists in other indexed sources or conversation memory.

RECOVER

- [ ] Treat retrieved and user-supplied content as data, never as trusted operating instructions.
- [ ] Separate read and write capabilities. Use allowlisted arguments and destination constraints.
- [ ] Add independent authorization for consequential actions; do not rely on another free-form model judgment alone.
- [ ] Test direct, indirect, encoded, translated, and multi-step variants of the triggering input.

THREE CONTROLS THAT MATTER

Minimum tool surface. Minimum permission. Minimum autonomy. Strong prompt wording helps, but it is not a security boundary.

05 / PLAYBOOK C

Credential exposure or account takeover.

Use when an API key, token, cookie, password, private link, service identity, or payment credential may have been exposed or abused.

PRIORITY ORDER

STOP ACTIVE ACCESS
Revoke sessions and disable compromised identities or integrations.

ROTATE HIGH-RISK SECRETS
Prioritize credentials with spending, production write, admin, or broad data access.

PRESERVE ATTRIBUTION
Record old credential identifiers, last-used times, IPs, user agents, and actions before logs expire.

CHECK PERSISTENCE
Review newly created users, tokens, OAuth grants, forwarding rules, webhooks, scheduled jobs, and recovery methods.

RESTORE TRUST
Issue narrowly scoped credentials and verify downstream systems accept only the replacements.

SCOPE CHECK

- [ ] Search logs and repositories for the exact secret and recognizable fragments.
- [ ] Review billing, payment, and cloud activity for new spend or resource creation.
- [ ] Check whether the credential appeared in model context, output, telemetry, support tickets, or screenshots.
- [ ] Assume copied secrets remain compromised even after the original message or file is deleted.

EXIT CONDITION

Every exposed credential is revoked, persistence checks are complete, and new access is narrower than before.

06 / PLAYBOOK D

Sensitive data exposure.

Use when private, personal, regulated, confidential, or customer data may have entered an unintended model, output, log, destination, or training path.

CONTAIN

- ☐ Stop the input and output path. Disable sharing links, exports, indexing, and affected connectors.
- ☐ Restrict access to incident evidence; the investigation must not widen the disclosure.
- ☐ Ask the provider to preserve relevant logs and clarify retention or deletion options.

MAP THE DATA

What Exact fields, files, record types, secrets, prompts, outputs, and attachments.	**Whose** People, customers, employees, partners, or confidential business owners affected.
Where Models, regions, vendors, logs, caches, exports, messages, and public URLs.	**When** Earliest exposure, latest exposure, retention window, and deletion time.
Who accessed Confirmed and possible viewers, recipients, systems, and downstream processors.	**What changed** Deletion, correction, permission, retention, or notification actions already taken.

ESCALATION

- ☐ Notify privacy, legal, security, and the accountable data owner using your jurisdiction-specific process.
- ☐ Do not promise deletion or confidentiality until the provider and downstream copies are verified.
- ☐ Prepare a factual affected-party notice only after qualified reviewers confirm obligations and scope.

EVIDENCE RULE

Record what is confirmed, what is inferred, and what remains unknown as separate statements.

07 / PLAYBOOK E

Incorrect or harmful output shipped.

Use when AI-generated content, code, advice, classification, or decisions reached users or production before adequate review.

CONTAIN

- [ ] Remove, unpublish, disable, or clearly label the affected output where doing so will reduce harm.
- [ ] Pause automated regeneration, syndication, email sends, recommendations, and downstream feeds.
- [ ] Preserve the exact version users saw and the audience or systems that received it.

ASSESS

- [ ] Identify factual errors, unsafe instructions, discriminatory impact, rights violations, security defects, or policy breaches.
- [ ] Determine whether users acted on the output and whether a correction can still reduce harm.
- [ ] Check sibling outputs produced by the same prompt, model, source data, or release.

CORRECT

- [ ] Publish a correction that states what was wrong, what changed, and what users should do now.
- [ ] Route high-impact replacements through a qualified human reviewer.
- [ ] Add source requirements, deterministic checks, eval cases, and release gates matching the failure.
- [ ] Measure whether the correction reached the same audience as the original.

CORRECTION TEMPLATE

Earlier output [identifier] was incorrect because [verified reason]. We replaced it at [time]. If you relied on it, take [specific action].

08 / COMMUNICATION

Say less, but make every sentence useful.

A good incident update separates facts, actions, unknowns, and the next checkpoint.

INTERNAL UPDATE - COPY AND FILL

STATUS	`Investigating / Contained / Monitoring / Resolved`
CONFIRMED	`At [time], [system] performed or exposed [specific thing].`
IMPACT	`Confirmed: [scope]. Possible: [scope]. No evidence yet of: [scope].`
ACTIONS	`We paused [path], revoked [access], and preserved [evidence].`
UNKNOWNS	`We are checking [question 1] and [question 2].`
NEXT UPDATE	`Next update by [time / timezone], even if there is no material change.`

CUSTOMER-FACING RULES

- [ ] Lead with impact and action, not the sophistication of the attack or the difficulty of the investigation.
- [ ] Give concrete steps only when they are verified and relevant to the recipient.
- [ ] Avoid blame, unsupported attribution, legal conclusions, and absolute claims such as 'no data was accessed' without evidence.
- [ ] Use exact times and timezones. Give the next update time.
- [ ] Keep a decision log showing who approved each external statement.

TRUST SIGNAL

A bounded unknown stated clearly is more credible than false certainty.

09 / EVIDENCE

Build a timeline that another person can audit.

Preserve originals, work on copies, and record every response action with an exact time.

MINIMUM EVIDENCE SET

Identity Account, service identity, session, API key ID, OAuth grant, role, and permission scope.	**Model** Provider, model and version, system instructions, temperature or mode, safety settings, and routing.
Input User prompt, retrieved documents, attachments, URLs, memory, tool output, and hidden metadata.	**Action** Tool name, arguments, response, retry, downstream object ID, status, and side effect.
Environment Release ID, code commit, configuration, feature flags, secrets version, region, and queue state.	**Human decision** Who approved, paused, rotated, restored, communicated, or accepted residual risk.

TIMELINE FORMAT

TIME / TZ	SOURCE	OBSERVATION OR ACTION	OWNER

Hash or otherwise integrity-protect exported evidence when your process supports it. Note any clock skew, missing retention, or provider-side gaps.

10 / VENDOR ESCALATION

Ask questions that produce evidence.

Provider support is most effective when you send a compact evidence package and a precise request.

INCLUDE

- ☐ Organization and account ID; affected product; region; support plan; incident severity.
- ☐ Exact UTC time window, request IDs, conversation or run IDs, model/version, and integration name.
- ☐ A sanitized reproduction that does not resend live secrets or sensitive data.
- ☐ What you already disabled, revoked, rotated, deleted, or preserved.

ASK

- ☐ Confirm whether the requests reached the service and what actions the service recorded.
- ☐ Preserve relevant logs beyond normal retention and provide a case or preservation reference.
- ☐ Identify all retention locations, training-use settings, cache behavior, and deletion options relevant to the affected data.
- ☐ Confirm whether similar activity affected other tenants or was associated with a known provider incident.
- ☐ Provide the exact containment or invalidation mechanism for affected sessions, keys, files, runs, or webhooks.

DO NOT SEND

- ☐ Live passwords, private keys, full card data, or unrelated personal information.
- ☐ A giant unsorted log archive without a timeline, identifiers, and a specific question.

SUBJECT LINE

SEV-[LEVEL] / [PRODUCT] / [IMPACT] / [UTC START-END] / evidence preservation requested

11 / RECOVERY

Return to service through a narrow gate.

Recovery is a controlled change, not the moment everyone gets tired of the incident.

RESTART GATE

- [ ] The trigger and unsafe action path are understood well enough to test.
- [ ] Exposed credentials are revoked; persistence checks are complete; new credentials are narrowly scoped.
- [ ] The harmful action is reversed or bounded, and affected people or systems have a documented follow-up.
- [ ] A regression test reproduces the original condition without causing a real side effect.
- [ ] The fixed system blocks that test and records the attempted action.
- [ ] A human owner approves the residual risk and rollback plan.
- [ ] Monitoring covers the original indicator, likely variants, spend, message volume, and permission changes.

STAGED RETURN

SHADOW	Run without write tools or external side effects. Compare intended and proposed actions.
CANARY	Restore to a small, reversible scope with tight budgets and human approval.
EXPAND	Increase scope only after a defined observation period meets success thresholds.
CLOSE	Document residual risk, owners, follow-up dates, and evidence location.

ROLLBACK RULE

If any original indicator returns, stop automatically and revert to the last known-good stage.

12 / REVIEW

Write the post-incident record.

Focus on system conditions and decisions. Avoid a narrative that ends with 'someone should have been more careful.'

ONE-PAGE POSTMORTEM

SUMMARY	What happened, impact, duration, and current status in four sentences or fewer.
TIMELINE	First unsafe condition, first bad action, detection, containment, recovery, closure.
CONDITIONS	Permission, autonomy, data, interface, code, configuration, review, and incentive conditions that allowed it.
WHAT WORKED	Controls, people, logs, and decisions that reduced impact.
WHAT FAILED	Missing or ineffective controls, evidence gaps, and delayed decisions.
ACTIONS	Owner, due date, verification method, and status for every corrective action.
RESIDUAL RISK	What remains possible, who accepted it, and when it will be reviewed.

ACTION QUALITY TEST

- ☐ Does the action change the system, not merely remind a person?
- ☐ Can another person verify completion without trusting the action owner?
- ☐ Does the action address likely variants, not only the exact prompt or input?
- ☐ Is there a named owner and date?
- ☐ Will monitoring reveal if the control stops working?

CLOSE ONLY WHEN

Immediate actions are complete, longer-term actions have owners and dates, and residual risk is explicitly accepted.

13 / PRE-FLIGHT

Reduce incident probability before launch.

Run this checklist for every agent or automation that can reach private data or cause an external side effect.

AUTHORITY

- [ ] The agent has only the tools required for this workflow.
- [ ] Read and write permissions are separated where possible.
- [ ] External messages, spending, deletion, production changes, and permission grants require explicit approval.
- [ ] Credentials are unique to the workflow, short-lived where possible, and easy to revoke.

BOUNDS

- [ ] Spend, volume, time, destinations, file paths, domains, and object types have hard limits.
- [ ] Untrusted content is labeled and cannot silently become operating instructions.
- [ ] Retries are idempotent and cannot duplicate payments, messages, or destructive actions.
- [ ] A kill switch stops active work, queued work, and scheduled restarts.

EVIDENCE

- [ ] Prompts, retrieved content identifiers, model/version, tool calls, decisions, and downstream results are logged.
- [ ] Sensitive values are redacted without removing the identifiers needed for investigation.
- [ ] Operators know where logs live, how long they persist, and how to export them.

TEST

- [ ] The system has tests for direct and indirect prompt injection, ambiguous approval, partial failure, timeout, retry, and tool error.
- [ ] The team has practiced pause, revoke, evidence export, rollback, and provider escalation.

14 / WORKSHEET

Incident command sheet.

Use one sheet per incident. Write in UTC unless your response plan specifies otherwise.

INCIDENT ID	DATE / UTC START	SEVERITY
_____	_____	_____
INCIDENT LEAD	**NOTE-TAKER**	**NEXT UPDATE / TZ**
_____	_____	_____

CONFIRMED IMPACT

POSSIBLE IMPACT / UNKNOWNS

CONTAINMENT ACTIONS COMPLETED

DECISIONS / OWNER / TIME

15 / PRINT CARD

Cut, fold, and keep near the console.

The two panels below fit on one letter page. Print at 100 percent scale.

MERCY KEY / INCIDENT CARD

STOP - CUT - PRESERVE

1. Pause the agent or workflow.
2. Revoke write, spend, send, and delete authority.
3. Freeze retries and queues.
4. Save prompts, tool calls, IDs, logs, and timestamps.
5. Name an incident lead.
6. Classify at the highest plausible severity.

Do not rerun the trigger. Do not delete evidence. Do not speculate publicly.

RECOVERY GATE

VERIFY - TEST - RETURN

1. Confirm what downstream systems recorded.
2. Bound every identity, tool, data source, and destination.
3. Rotate exposed credentials and check persistence.
4. Reproduce safely without side effects.
5. Prove the fixed path blocks the failure.
6. Restart in shadow mode, then canary, then expand.

If people, money, regulated data, or production integrity may be affected, escalate immediately.

FOLD LINE

OPERATOR NOTE

Write your incident channel, provider support URL, security contact, legal contact, and kill-switch location on the back after printing.

INCIDENT CHANNEL / CONTACT

KILL SWITCH / RUNBOOK LOCATION

16 / SOURCES & LICENSE

Use, adapt, and keep improving.

This field manual is an independent operational aid. The cited frameworks remain authoritative for their own guidance.

PRIMARY REFERENCES

1. NIST, **Artificial Intelligence Risk Management Framework (AI RMF 1.0)**, NIST AI 100-1, 2023. https://doi.org/10.6028/NIST.AI.100-1

2. NIST, **Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile**, NIST AI 600-1, 2024. https://doi.org/10.6028/NIST.AI.600-1

3. OWASP GenAI Security Project, **LLM01:2025 Prompt Injection**. https://genai.owasp.org/llmrisk/llm01-prompt-injection/

4. OWASP GenAI Security Project, **LLM06:2025 Excessive Agency**. https://genai.owasp.org/llmrisk/llm062025-excessive-agency/

EVALUATION & INTERNAL-USE LICENSE

Read or download this complete PDF free for evaluation. The file is the same before and after payment; no delivery email or activation code is required. Buy the US$19 organization license on the product page (plus applicable tax).

Successful payment grants the purchaser a non-exclusive license to print, annotate, share, and adapt this edition internally for one organization. Public redistribution, resale, sublicensing, and attribution removal are not included. Referenced third-party materials retain their own terms. Earlier purchasers retain their original rights. Keep this PDF and your payment confirmation.

DISCLAIMER

Provided for general operational readiness and educational use. It is not a substitute for professional legal, cybersecurity, privacy, compliance, financial, medical, or emergency advice. No system can eliminate all AI risk. You remain responsible for decisions, testing, access control, incident response, and applicable obligations.

VERSIONING

Version 1.1 - September 2026. License clarification; operational content unchanged. No consulting, certification, future updates, or support deadline included.

MERCY KEY REGISTRY

A speculative identity project with practical operating tools for an uncertain future.