Rehearse an AI incident

24 evidence-led exercises | Content v0.4 | 7 November 2026

QUICK START: Choose one exercise, not the whole library. Allow 90 minutes. Read this page and the timing page; the facilitator reads that exercise's private pages before play. Use made-up data only. No real operational action is authorised by this pack.

What you leave with

A bounded account of what happened, decisions with owners and evidence limits, words you can adapt for a real escalation, a tested-on-paper containment plan, a prioritised fix plan and a first-ten-minutes card. These are rehearsal outputs, not assurance that controls work.

What you need

Paper, pens and a timer. Use these two common pages and the eight-page exercise. Keep facilitator pages private until their release point. The pack works offline with no account, form, email submission or network connection. Completed sheets stay locally under your normal record rules. No real secrets, personal records or private account credentials belong on these sheets.

Opening words

"This is a rehearsal. Every exhibit is made up. We will decide using what is supplied, not facts we wish existed. State a decision, owner and evidence check. We can act under uncertainty, but we cannot make uncertainty disappear. Nobody is being graded. If we discover a live incident, we stop play and use our real procedures."

Use the right authority

No legal advice, compliance conclusion, safety certificate or universal technical operating instruction is supplied. Use qualified people and your established procedures for specialist decisions. Communication examples are rehearsal drafts, not approval to send. A workshop cannot authorise a real payment, access change, notification or customer remedy.

Pick a mode

3-6: facilitator plus exercise roles; combine hats as shown in each kit. Two: A facilitates and holds evidence, B holds the other roles, then both challenge decisions. Solo: cover later exhibits, injects and facilitator notes; write each answer before revealing the next. An editorial walkthrough is possible alone, but it is not a live team test.

© Almost Magic Tech Lab. For your organisation's own use only. Please do not copy, share or sell this material.

Run one exercise in 90 minutes

Time	Facilitator action
Before	Read private pages 3-6 of the kit. Mark the two branch consequences you may need. Choose a recorder. Do not invent additional technical facts.
00-10	Read the opening and roles. Release exhibit A. Ask players to state the operational decision before collecting more information.
10-20	Read inject 1. Release exhibit B. Record a decision, owner and what would disprove it.
20-30	Read inject 2. Release exhibit C. If a listed branch matches the decision, read its consequence as a new fictional event. Otherwise keep the base timeline.
30-40	Read inject 3. Rehearse the holding statement and escalation aloud. Test the team's assumed stop boundary.
40-50	Read inject 4. State the final unknown. Do not supply a missing record to rescue the team.
50-65	Use the four insight questions and strong/weak/misleading answer guide. Ask players to amend one decision after a counterexample.
65-80	Complete the evidence, control and decision sheets. Pick the top three fixes; state what would prove each worked.
80-90	Debrief, complete the three-label scorecard and keep card, and set the day-90 comparison.

Branch rule: consequences are conditional rehearsal events, not predictions about a real system. They expose a trade-off. Do not call every broad restriction wrong or every narrow restriction right. Accept alternatives that name the authority, side effects, missing evidence and a falsifiable check.

The only scorecard labels are "we decided", "we recorded", "we could not confirm". No grades, percentages or certificates. At 90 days compare evidence with this run, not confidence. If a new scenario is used, do not claim like-for-like improvement.

© Almost Magic Tech Lab. For your organisation's own use only. Please do not copy, share or sell this material.

11 | The dataset that taught the wrong rule

1 / 8 | PARTICIPANT BRIEF AND ROLE CARDS | All records, numbers and quotations are fictional.

Training-data contamination - Distinguish a changed model behaviour from a bad evaluation or a bad data lineage.

Read aloud

"A newly tuned internal classifier begins routing urgent requests into a low-priority queue. The release used a freshly labelled training dataset assembled from several teams. The previous model did not show this pattern in its recorded checks."

Decision to bring to the table

Within the first discussion, choose the interim operating boundary and name its decision owner. Separate what stops further impact from what would establish past impact. You may choose a restriction while facts remain missing; record its side effects and next review.

Decision lead

Choose a temporary routing method and release restriction. State what you can decide in this rehearsal and what needs another authorised person. Do not assume a job title creates legal authority.

Data lead

Trace labels, sources and dataset versions. State what you can decide in this rehearsal and what needs another authorised person. Do not assume a job title creates legal authority.

Operations lead

Find delayed urgent work and organise human triage. State what you can decide in this rehearsal and what needs another authorised person. Do not assume a job title creates legal authority.

Evidence lead

Compare evaluation samples with the observed failures. State what you can decide in this rehearsal and what needs another authorised person. Do not assume a job title creates legal authority.

Modes

3 people: facilitator holds evidence, one player is decision lead and one combines the two other operational roles. 4: facilitator holds evidence, each player holds one remaining role. 5: separate facilitator and four roles. 6: add an observer who records unsupported assumptions, not a rescuer who invents facts.

Two: A facilitates and holds evidence. B holds the other three roles and announces the current hat. A reveals only the current exhibit/inject. B writes one decision before the next reveal. Both challenge a decision during debrief. Keep the 90-minute clock.

Solo: cover pages 3-6 and later exhibit sections. Read the opening and exhibit A, write a decision, then reveal each inject and its exhibit. Answer as decision lead, operator and affected person separately. Only then uncover facilitator calibration. Flag every success depending on invented facts.

Do not browse, contact anyone or operate a real system during play. If an exercise raises a live concern, stop the rehearsal and use the established incident route.

© Almost Magic Tech Lab. For your organisation's own use only. Please do not copy, share or sell this material.

11 | The dataset that taught the wrong rule

2 / 8 | CONTROLLED-RELEASE EVIDENCE DOSSIER | All records, numbers and quotations are fictional.

A | Routing results

Release at opening. Old model routes 18 of 20 urgent test requests correctly. New model routes 11 of 20. Review calls the new evaluation green because it uses a different 12-case set. The model version changed; queue rules did not.

B | Dataset lineage

Release with inject 1. Training batch B6 contains 240 examples. A sampled urgent case is labelled routine. Forty source rows have no provenance. The evaluation set includes near-duplicates of training examples; no independent challenge set is attached.

C | Service state

Release with inject 2. Rollback returns old predictions on new requests. Fifty-two previously routed items remain in the routine queue. Staff begin relabelling those items, but do not preserve the original model output. One source batch cannot be reconstructed.

Evidence challenge before discussion

For each exhibit name: the source, what it directly supports, what it does not support, whether its time boundary matches the incident, and what independent corroboration would matter. A screenshot, recollection or dashboard may be useful without being complete.

Record	Direct observation vs inference / missing test
A	
B	
C	

Facilitator handling: later facts may change a decision, but they do not retroactively improve the evidence available when it was made. Record revisions; do not erase the earlier reason. No exhibit establishes the final missing fact on page 3.

© Almost Magic Tech Lab. For your organisation's own use only. Please do not copy, share or sell this material.

11 | The dataset that taught the wrong rule

3 / 8 | FACILITATOR: INJECTS AND BRANCH CONSEQUENCES | All records, numbers and quotations are fictional.

Inject 1 | 10-20 minutes

Say: "A small sample shows urgent examples labelled routine. Nobody signed off that label batch."

Probe: Ask for lineage and review responsibility, not a quick blame assignment.

Inject 2 | 20-30 minutes

Say: "The release dashboard is green because its test set came from the same label batch."

Probe: Identify leakage between training and evaluation. Require an independent challenge set.

Inject 3 | 30-40 minutes

Say: "Rolling back restores older behaviour but leaves the misrouted queue untouched."

Probe: Treat the service impact separately from the model rollback.

Inject 4 | 40-50 minutes

Say: "One source batch lacks provenance and cannot be reconstructed. Can the team prove which training records caused each error?"

Probe: Individual causation cannot be established from the pack. Record the missing lineage and define a bounded retraining decision.

Two conditional branches

1. If players choose: Roll back and declare all impact reversed. Then say: New predictions improve in play, but the old queue stays misrouted. Containment must include already-made decisions and time-sensitive work.

2. If players choose: Retrain immediately on corrected queue labels. Then say: Some corrections reflect later human outcomes rather than original urgency. Without lineage and independent evaluation, the next model can conceal the same failure.

After either branch require an amended decision, not a promise to be more careful. Ask which side effect they accept, who can authorise it, how it is reversed and which observation would trigger a change. If their choice is not listed, do not fabricate success or harm; test the reasoning against the supplied exhibits.

© Almost Magic Tech Lab. For your organisation's own use only. Please do not copy, share or sell this material.

11 | The dataset that taught the wrong rule

4 / 8 | FACILITATOR: INSIGHT QUESTIONS AND ANSWER CALIBRATION | All records, numbers and quotations are fictional.

Which comparison is valid when evaluation sets differ?

Strong reasoning to listen for: Run both versions on a fixed, independent set with relevant strata. Do not compare green labels from different samples as if they measure the same thing.

What must be retained before staff correct the queue?

Strong reasoning to listen for: Original output, version, request time, correction and basis. Otherwise impact review becomes inseparable from repair.

Can a bad source batch explain each error?

Strong reasoning to listen for: Not from the supplied lineage. It is a plausible contributor, not proven individual causation.

What second-order harm follows restoring the old model?

Strong reasoning to listen for: Affected queued work can still age, and the old model may have known limits. Choose continuity and backlog reconciliation, not a ceremonial rollback.

Weak is incomplete; misleading is unsupported

Weak answer: "Roll back and correct the training labels."

Why to probe: this names an activity but leaves the mechanism, scope or proof unfinished. Do not accept activity as an outcome.

Misleading answer: "Both evaluation reports are green, so the new model is at least as good."

Next probe: Ask whether the same independent cases, labels and denominators were used, and what happens to the old queue.

A strong team need not choose one prescribed solution. Require a bounded decision, an accountable owner, a reversible path where feasible, a specific observation that would falsify success, and explicit missing evidence. If a team gives a different well-supported answer, capture it rather than forcing the model answer.

Final-inject guardrail: Individual causation cannot be established from the pack. Record the missing lineage and define a bounded retraining decision. Do not turn an expert-sounding explanation into a new exhibit.

© Almost Magic Tech Lab. For your organisation's own use only. Please do not copy, share or sell this material.

11 | The dataset that taught the wrong rule

5 / 8 | READY-TO-USE PRESERVATION AND COMMUNICATION DRAFTS | All records, numbers and quotations are fictional.

Evidence preservation request

"Please preserve the following in its authorised location without changing the original: Preserve dataset and model versions, sample labels, split definitions, original routing outputs and queue snapshots. Keep corrections as new records rather than overwriting original evidence. Record who collected it, when with timezone, source/version, access restriction and any known gaps. Use a restricted reference rather than copying sensitive content into the incident chat."

Item / source reference	Collector / time / scope	Integrity or completeness limit

Holding statement - adapt, then approve

"We have restricted the new routing model and are reviewing work it already classified. Reverting the model does not automatically correct the queue. Missing source lineage prevents us from attributing every error to a particular training record."

Before any real use: recipient/audience _____ approved decision owner _____ next update time/channel _____. Verify every factual clause against the current record. Do not imply the restraint or fallback was completed if it is only proposed.

Escalation script

"Please assign separate owners for model rollback and backlog recovery. The evaluation sets are not comparable and some original outputs are being overwritten. We need a preserved baseline and a fixed independent challenge before another release."

Recipient function / authorised route: _____

Decision needed by / fallback if unanswered: _____

Contradiction check

Circle any phrase your evidence cannot support. Read the statement as the affected person: would they infer complete containment, no impact, a guaranteed remedy or a confirmed cause? Amend it. Name who will correct it if later evidence changes the picture.

© Almost Magic Tech Lab. For your organisation's own use only. Please do not copy, share or sell this material.

11 | The dataset that taught the wrong rule

6 / 8 | KILL-SWITCH, OWNERSHIP AND SUPPLIER WORKSHEET | All records, numbers and quotations are fictional.

Exercise-specific containment boundary

Put urgent work through a named human triage path; stop new model routing; reconcile already-routed items. Verify both new-request behaviour and backlog handling.

Control question	Named owner / evidence / limit
Who authorises the stop, who executes it and who verifies it?	
Which worker, queue, cache, identity or human action can bypass it?	
What legitimate work is interrupted; who owns the fallback?	
What observation would prove the stop failed?	
What are the restart conditions and who may accept residual unknowns?	

Supplier / internal system-owner questions

1. Can model, dataset and evaluation versions be linked?

Answer / record reference / absent evidence: _____

2. Which records crossed training/evaluation boundaries?

Answer / record reference / absent evidence: _____

3. What source provenance was never retained and cannot be recovered?

Answer / record reference / absent evidence: _____

Do not accept "resolved", "equivalent", "deleted", "successful" or "all tests passed" without the definition, scope and supporting observation. Ask the supplier or system owner to state what was not measured. A missing response is an evidence gap, not proof of the opposite.

Request owner / due date / escalation route: _____

What can be decided before this answer arrives: _____

© Almost Magic Tech Lab. For your organisation's own use only. Please do not copy, share or sell this material.

11 | The dataset that taught the wrong rule

7 / 8 | CAPTURE SHEET AND BRANCH DECISION LOG | All records, numbers and quotations are fictional.

Session date / mode: _____ participants by role: _____

Inject / exhibit	Known fact and exact source	Assumption / missing evidence	Decision / owner / check
1			
2			
3			
4			

Decision record - revise, do not erase

Decision / time / authority	Rejected alternative / side effect	Success test / falsifier	Review time / owner

Branch taken / consequence / changed decision: _____

_____

Final unknown in our own words: _____

What we do while it remains unknown / who can revisit: _____

_____

© Almost Magic Tech Lab. For your organisation's own use only. Please do not copy, share or sell this material.

11 | The dataset that taught the wrong rule

8 / 8 | PRIORITISED FIX PLAN, DEBRIEF AND KEEP CARD | All records, numbers and quotations are fictional.

Proposed priorities - assign people and dates locally

Priority / accountable function	Fix and evidence that would prove it worked
Now Operations owner	Reconcile urgent backlog with original outputs retained.
Within 7 days Model owner	Run old and new versions on the same independent challenge set and record category-level failures.
By day 30 Data owner	Demonstrate source-to-label lineage and a held-out set that rejects near-duplicate leakage.

Named owners / actual dates / record location: _____

Debrief and three-label record

Did a green evaluation become a substitute for source quality?

Were old queue effects recovered as well as new behaviour?

Only these record labels	Specific evidence or explicit limit
we decided	
we recorded	
we could not confirm	

Day-90 recheck

Repeat date _____. Keep the original decision log and evidence references. Re-run this scenario with the same critical test, then introduce the branch you avoided. Compare the top-priority proof above with the first run. Record changed scope or conditions; a different denominator is not improvement. State one unresolved unknown even if all fixes are marked done.

First ten minutes - detach and keep

Restrict the new model's routing role and put urgent work through a named human triage path.

Decision owner / authorised route _____ containment owner _____

Evidence reference / keeper _____ human fallback / specialist _____

Still unknown / next check time _____

Rehearsal prompt only. Use established procedures and qualified help in a real incident. Do not keep sensitive details or secrets on this card.

© Almost Magic Tech Lab. For your organisation's own use only. Please do not copy, share or sell this material.