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.

04 | The invoice hidden in the document

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

Indirect instruction attack - Keep document content from becoming authority to change a payment.

Read aloud

"An AI assistant summarising supplier attachments produces a request to change bank details. The request came from text inside a PDF, not from the person authorised to approve payments. No payment has yet been made."

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 the payment hold and escalation path. State what you can decide in this rehearsal and what needs another authorised person. Do not assume a job title creates legal authority.

Workflow lead

Isolate the document-processing route without deleting the evidence. State what you can decide in this rehearsal and what needs another authorised person. Do not assume a job title creates legal authority.

Accounts lead

Verify the invoice through an independently held contact route. 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

Track which statements came from the file and which from trusted approval. 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.

04 | The invoice hidden in the document

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

A | Invoice intake timeline

Release at opening. 09:02 supplier attachment received. 09:04 assistant summary proposes changing beneficiary details. 09:05 human clicks approve summary. 09:07 payment draft created; no payment released. Original supplier contact record has a different account reference.

B | Extracted document text

Release with inject 1. Visible invoice body requests payment for routine work. A separate extraction region contains a note asserting that the change is already authorised. The summary quotes that assertion as approval. This exercise does not supply an attack payload or executable instructions.

C | Approval and contact history

Release with inject 2. Approval event references summary version S2, not the uploaded document hash. A callback was made to the number in the attachment. The independently stored supplier contact has not been called. Processing log loses events between extraction and summary generation.

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.

04 | The invoice hidden in the document

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

Inject 1 | 10-20 minutes

Say: "The assistant calls the change already approved. Its only citation is the supplier attachment."

Probe: Make the team identify the actual authority record. A claim of approval inside a file is not approval.

Inject 2 | 20-30 minutes

Say: "The payment batch closes soon. A colleague wants to accept the change because the invoice amount is familiar."

Probe: Test whether urgency or a correct amount substitutes for destination verification. Keep the payment decision with its owner.

Inject 3 | 30-40 minutes

Say: "The supplier answers using the address in the suspicious attachment and confirms the new details."

Probe: Ask whether the confirmation route was independent. Do not provide operational attack text; rehearse verification only.

Inject 4 | 40-50 minutes

Say: "The original upload source and a portion of the processing log are missing. Was this deliberate interference or an accidental note?"

Probe: Intent cannot be established from the pack. Record the observed behaviour and uncertainty without attributing an attacker.

Two conditional branches

1. If players choose: Treat the human approval click as the control working. Then say: The click approved a summary carrying unverified authority. The payment draft advances in play. Ask what the reviewer actually saw and which independent fact the click established.

2. If players choose: Delete the attachment and restart intake. Then say: The payment remains held, but the original source and transformation evidence are lost in play. Containment and evidence preservation need different actions and owners.

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.

04 | The invoice hidden in the document

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

Where exactly did factual content become decision authority?

Strong reasoning to listen for: At the summary claim and approval workflow, not merely at document upload. Trace the asserted authorisation to an independently held approval record.

Can a trusted supplier document still contain untrusted workflow instructions?

Strong reasoning to listen for: Yes within this fictional workflow. Authentic document origin would not make every embedded request an approval under the organisation's process.

What does a successful callback prove if the number came from the disputed document?

Strong reasoning to listen for: Only that someone answered that number. It is not an independent destination check.

How would you test the repair without building a better attacker?

Strong reasoning to listen for: Use a harmless made-up document that makes an unauthorised approval claim. Show the system treats it as quoted content and holds the destination change.

Weak is incomplete; misleading is unsupported

Weak answer: "Train staff to read invoice summaries more carefully."

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

Misleading answer: "The approver clicked yes, so the destination change was authorised."

Next probe: Ask what independent approval was displayed, not whether a click was logged.

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: Intent cannot be established from the pack. Record the observed behaviour and uncertainty without attributing an attacker. 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.

04 | The invoice hidden in the document

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 original bytes, upload identifier, extraction regions, summary versions, what the approver saw, approval event and payment draft. Restrict access; do not forward the document to every reviewer. 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

"A proposed payment-detail change is on hold while we verify it through our established contact and approval route. No payment has been released in the supplied record. We have not established why the unauthorised instruction appeared."

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

"The summary converted an attachment assertion into approval evidence. Please keep the payment held, assign an accounts verification owner and preserve intake-to-approval versions. We need an independently sourced contact, not another reply through the disputed attachment."

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.

04 | The invoice hidden in the document

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

Exercise-specific containment boundary

Hold beneficiary edits and payment release separately. Isolate document-derived actions while keeping read-only evidence. Verify the held draft cannot be released by another queue or approver.

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 you provide original-to-summary provenance including extraction regions?

Answer / record reference / absent evidence: _____

2. Which workflow fields can document content influence?

Answer / record reference / absent evidence: _____

3. What approval record is independent of the document and visible to the final reviewer?

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.

04 | The invoice hidden in the document

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.

04 | The invoice hidden in the document

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 Accounts owner	Confirm draft and release holds through an independent ledger view.
Within 7 days Workflow owner	Demonstrate a document claim cannot satisfy approval or change a beneficiary.
By day 30 Control owner	Run a reviewer test showing original facts, quoted claims and trusted approval are visibly separate.

Named owners / actual dates / record location: _____

Debrief and three-label record

Did the team preserve provenance through the summary?

Did a deadline change the approval standard?

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

Hold the affected payment change and isolate the document-driven action route; use an independently verified accounts contact.

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.