

CASE 14/30

Investigating an AI service failure: two companies, one agent, one error

Sergei Ponomarev · AI Business · September 2026

Everyone who runs an AI agent wants to be sure of one thing: that the agent does exactly what it is supposed to do, and does not do what it is not supposed to do.

The hard part is how you make sure. Not how you prevent failure, because you cannot prevent all of it. An agent drifts. A supplier ships an update. Behaviour changes quietly, and nobody notices until a customer does. The real question is how you build a system of checks that lets you find the failure fast, establish what actually happened, and limit the damage.

I am the author of a package built for exactly that: a four-part system comprising three operational documents, the AI Policy, the AI Service Passport and the AI Receipt, together with the Rules governing their implementation. Let me be plain about what it does and does not do. **The system does not prevent errors.** If an agent drifts, it drifts. What it does is make the failure findable, the norm provable and the consequences traceable, so that the company can react within days instead of arguing with itself for weeks.

This document is a thought experiment that shows the difference. Two identical companies suffer the same failure at the same time from the same cause. One has the system implemented, the other does not. Everything that follows is what each of them can and cannot establish afterwards, and how much of the harm each one can actually undo.

The companies and all circumstances are fictional. This is a modelled situation, not a report on any real company. Numerical values are given for illustration; quantifying the saving requires measurement in a controlled simulation or in a real incident.

1. Conditions of the analysis

The same failure is examined in two companies that are alike in service, scale, base agent and the nature of the failure. There is one difference between them: whether an implemented system of requirements, records and controls was in place beforehand.

Alpha LLC and **Omega LLC** are fictional online retailers of household goods. Comparable range, around ten thousand orders a month, a 30-day returns rule. Both use the same base AI agent from the same supplier: it answers product questions, places orders and returns, reports delivery status and takes complaints.

Omega uses the standard product. The agent runs in its default configuration and supplier updates are applied automatically. Its means are a system prompt, the platform dialogue log, an order management system, a customer database and a support mailbox. This is common good faith practice, not a caricature.

Alpha commissioned the same supplier to adapt the agent to its system. The agent works from the organisation AI Policy and the AI Service Passport as its sources of rules, uses only the sources of information the passport provides for, issues an AI Receipt for every interaction recording the criteria applied and the information used, and supplier updates are entered in a change register and checked before they are applied. For the comparison to stay honest: that pre-check did not include the scenario of a return request between day fifteen and day thirty, so the defect passed the check at Alpha too.

For the purposes of the experiment, Alpha operates a four-part system: a public AI Policy of the organisation; an AI Service Passport, the standard for the specific service, with annexes; the AI Receipt, an interaction record automatically generated for every started interaction, with a customer-facing part and a machine-readable technical record; and the Rules governing their implementation. Provisions are referred

to by the content of their sections, without clause numbers.

On the cause. For the designer of the thought experiment, the common underlying cause is stipulated to be a supplier update. Neither company is assumed to know or be able to prove this at the moment of discovery. What follows is written from the position of the investigators, not of the author.

One material clarification. The advantage is created by an implemented system of requirements, records and controls, not by the existence of documents. Four files prove nothing on their own. The conclusions here relate to the presence of versioning of the norm, the recording of interactions, the registration of changes, traceability of outputs and the allocation of responsibility. An equivalent system under different names is capable of producing a comparable result.

2. Established facts of the failure

For some period the agent at both companies told customers the return period was fourteen days instead of thirty, and automatically closed requests received after day fourteen as out of time. The failure was discovered in the same way at both companies: a customer who had been refused disputed the charge with their bank and left a public review.

At the moment of discovery, the following was established at each company:

- no fewer than forty customers were refused a return for this reason, and each was sent a refusal message;
- some customers left negative reviews, some went to their banks to dispute payments;
- the warehouse, not receiving the expected returns, revised its stock forecast;
- marketing attributed the drop in returns to the new product packaging;
- it is not known what other decisions may have been taken on the basis of this data;
- the exact time at which the defect appeared is not known.

The developer fixed the agent, and a repeat check of the same scenario now returns thirty days. A successful repeat check is not accepted as proof that all consequences have been removed. The figure of forty shows the number of refusals identified; it does not prove the completeness of the population of affected customers. A negative review following a refusal does not prove a causal link until the specific customer, the interaction and the complaint have been matched to one another.

3. The investigation at Alpha

Containment before the fix.

Until the initial check is complete, the automatic refusal function is suspended, return requests are passed to staff, and the original records are preserved unaltered. The failure is given an incident identifier, and all affected records, systems and measures are linked to it.

The norm.

The section of the passport on time limits and grounds for refusal fixes the 30-day rule, and every AI Receipt carries the versions of the AI Policy and the passport that were in force. The declared wording in force is established unambiguously for any date. The investigation begins by comparing the norm with actual behaviour, not by searching for the norm.

Boundaries and cause.

The change register gives a limited list of candidate causes, and the candidate closest to the first errors is a supplier update to the agent. A reverse search is run over the machine-readable receipts for every interaction where the value of fourteen days was stated or a refusal with the status out of time was generated. Matching the first erroneous record, the last confirmed correct one and the date of the update narrows the window. The configuration or deployment identifier recorded in the receipts makes it possible to test the link between the failure and a specific update. Where the supplier keeps no stable identifier, the update remains a hypothesis as to cause. Residual uncertainty is not concealed but stated explicitly.

Affected customers and completeness of records.

Those affected are identified by selecting on the receipts: service, return procedure, period, criterion applied, refusal status. The completeness of the selection is tested separately: the register of interactions and the system events are reconciled against the number of receipts issued, and a missing expected receipt is treated as a deviation in its own right. The result is a reproducible list within the recorded population, with completeness tested against the specified control totals. An event that was never recorded anywhere remains outside it.

Data used.

The operations log of each receipt records the criteria applied and their values, together with the information actually used. What can be proved is therefore not only that the agent answered incorrectly, but that the fourteen day criterion was applied in the decision, and in which interactions exactly.

Recipients of the erroneous output.

An annex to the passport contains a map of the outputs of the service: what result or status is produced, where it is recorded, who receives it directly and who uses it downstream. For this failure the outputs are the message to the customer, the status of the request, the returns statistics and its registered consumers, the warehouse and marketing. The investigation follows each output in turn.

Remediation in three layers.

Remediation is split into three tasks, and substituting one for another is not permitted: the behaviour of the agent (thirty days again); the position of the affected customers (refusals are reviewed by staff, results are recorded as linked entries, and the original receipts are not rewritten); and the consequences in adjacent systems (tasks to correct the warehouse forecast and the marketing conclusions are passed to the owners of those processes, and completion is recorded in the incident file). For each affected interaction, the measure, the responsible person, the deadline and the state of completion are recorded.

Resumption and lessons.

A successful repeat check remains evidence, not permission. The decision to resume the service is taken by the person accountable for the use of AI, as defined by the AI Policy, and the grounds are recorded. Following the incident the check programme itself is revised: the scenario of a request after day fourteen is added to the test purchases, because the failure showed the test coverage to be insufficient. The recorded agent versions and the change register make it possible to single out the supplier update as a testable hypothesis of cause and to assemble evidence for its technical and contractual assessment.

4. The investigation at Omega

Omega may well possess backups, platform logs or supplier release notes. The case assumes only that these materials were not maintained as a connected, versioned and reproducible evidence system.

The norm.

The 30-day rule exists in several texts: on the website, in the terms of sale, in the support handbook and in the system prompt of the agent. Which wording was in force on the date of each refusal cannot be established from the records available, because no edit history was kept and no versions of the texts were archived. The first dispute in the investigation is therefore not about the failure at all. It is about what counted as the norm on a given day.

Boundaries and cause.

The agent update was applied automatically and is recorded nowhere on the company side. The date on which the failure began is unknown, and the dialogue log is not linked to agent versions. The search is conducted through manual exports on the string fourteen days. A conservative boundary is applied: from the last day on which the service is reliably known to have worked correctly, to the date of the fix. The window is knowingly wider than the real one, and the link between the failure and the supplier update remains a supposition with nothing to confirm it.

Affected customers.

Those affected are sought by manually matching dialogue exports, closed requests, reviews and bank disputes. There is nothing with which to confirm completeness: no register of interactions exists, and the company cannot establish that the available dialogue records are complete for the relevant period, since some refusals exist only in email. The forty refusals found are what was identified, not what is complete.

Recipients of the erroneous output.

There is no pre-existing list of the consumers of returns data. That the erroneous statistics went into the warehouse forecast and the marketing conclusions comes out through asking departments. Any unregistered report or export may remain outside the investigation entirely.

Remediation and resumption.

The supplier confirms that the agent has been fixed. Who is authorised to decide that the service may continue was not determined in advance, so the decision is taken as the situation demands and the grounds are not recorded. A claim against the supplier is difficult, because the company cannot prove which version of the agent was in force at the moment of each refusal.

Result: Omega is able to fix the agent, but a substantial part of the investigation goes on reconstructing a missing control environment after the event, and a number of key questions cannot be reliably answered from the records available.

5. Twelve questions any such investigation has to answer

Answers are given separately for the two modes: Alpha, with the system implemented, and Omega, without it.

Question	Alpha	Omega
1. What is an established fact and what is unknown	an established fact is supported by a verifiable record or a set of consistent evidence; absence of support means the circumstance is not established	the line between fact and supposition is contestable; the completeness of the body of data cannot be proved

Question	Alpha	Omega
2. What the repeat check proved	current behaviour in the scenario checked; it forms part of the body of supporting evidence but does not replace it	the same narrow proof, but there is no body of supporting evidence
3. What further evidence is needed	determined from carriers known in advance: receipts, change register, output map, archive of wordings	the list of what is missing is itself a subject of the investigation
4. How to determine the defect window	the change register gives candidates; a reverse search over receipts narrows the window; residual uncertainty is recorded	a conservative boundary from the last reliably correct day; the window is knowingly wider than the real one
5. Who consumed the incorrect output	from the output map of the service and its registered consumers: customer, accounting, warehouse, marketing	by asking departments; unregistered consumers may be missed
6. Which kinds of consequence are separated	three layers: behaviour of the agent; position of affected customers; consequences in adjacent systems	the layers merge, and "the bot is fixed" substitutes for removing the consequences
7. Authority and accountability at each level	the service owner under the passport; the person accountable for the use of AI under the AI Policy; owners of adjacent processes receive tasks that are recorded	roles are not defined in advance and are allocated as the situation demands
8. The boundary of the AI service owner	the service owner answers for the service; erroneous outputs beyond its boundary are passed to process owners through the output map	the boundary is not set; the question is settled when a dispute arises
9. Conditions for resuming the local service	containment, preservation of evidence, removal of the cause, corrective measures, repeat check, absence of regression	there are no conditions established in advance
10. Conditions for resuming the wider process	confirmation of the fix for each output and each registered consumer; the decision is taken by the owner of that process or another authorised manager	determined as the situation demands, if they are posed at all
11. Who takes the decisions to resume operation	two separate decisions: resuming the specific AI service is authorised by the person accountable for the use of AI; resuming the wider process is authorised by its owner or another authorised manager; the grounds for both are recorded	not defined; the decision may remain undocumented
12. What the approach cannot do by its own means	see section 7: the limits are set out explicitly	the limitation is not explicitly recognised; the unknown is indistinguishable from the un-investigated

6. Summary comparison

Investigation question	Alpha	Omega
Which norm was in force	established from the passport version and the archive of wordings	reconstructed from texts that do not match
What happened in the interaction	from the AI Receipt and the technical record	reading dialogues, manual matching
When the failure began and what caused it	window narrowed; supplier update as a testable hypothesis	date unknown; wide estimated boundary

Investigation question	Alpha	Omega
Who was affected	reproducible selection within the recorded population	manual search; completeness cannot be proved
Are the records themselves complete	register of interactions reconciled against the specified control totals	nothing to reconcile against
What data the agent used	criteria and information recorded	cannot be reliably established
Where the erroneous data went	by the output map and its registered consumers	by asking around, with a risk of omission
Has the customers' position been put right	measure, responsible person and deadline for each	there is no single object of control
Who decides on resumption	defined, and the grounds are recorded	not defined
What to show an auditor or an insurer	a linked package of records with versions	exports and explanations after the event

The formula for the comparison: without a connected system of requirements, records and controls, an organisation has to reconstruct both the incident and its own control environment after the event. With such a system, it investigates the incident against a versioned map prepared in advance.

7. What this analysis does not establish

- The system did not prevent the failure. It happened at both companies at the same time and from the same cause, and this case does not hide that.
- Having documents does not in itself prove that requirements were complied with. In the Alpha mode it is assumed that the requirements are actually implemented in the service and that records are generated from real events.
- The passport identifies the recorded organisational rule that the agent was instructed to apply. The analysis does not establish that this rule is legally valid or that it overrides applicable law or binding contractual terms.
- The analysis does not establish legal liability, insurance cover or the size of any loss.
- A causal link between an individual refusal and a subsequent review or bank dispute requires separate confirmation in each case.
- Information that was never recorded is not recoverable by any system. Completeness can be tested within the recorded population, never beyond it.
- The completeness of the output map depends on it being maintained as the service changes. That is a condition, not a given.
- The quantitative saving in time and cost has not been measured and requires a controlled simulation or a real incident.

8. Testable hypotheses

The thought experiment becomes measurable if the following quantities are compared between the two modes. Two of them can be measured against ground truth only in a controlled simulation in which the evaluator knows the actual defect window and affected population but withholds them from the investigation teams. In a real incident those true values are normally unknown, and only the remaining measures apply.

Measured against ground truth, simulation only:

- the width of the established defect window relative to the actual one;
- the proportion of affected interactions found, against the actual number;
- the proportion of interactions wrongly included in the affected population.

Measurable in a real incident as well:

- the time taken to establish the norm in force on a given date;
- the time to containment;
- the time to a justified decision to resume operation;
- the number of manual sources that had to be reconciled;
- the completeness with which recipients of erroneous data are identified;
- the working hours expended on the investigation;
- the number of questions left without a proven answer;
- the time taken to assemble an evidence package for an auditor or an insurer.

Annex. Controls and records of the two companies

Controls and records	Alpha	Omega
AI agent	the same agent, adapted: issues AI Receipts, records criteria and information	standard product, default configuration
Agent updates	registered and checked before they are applied	applied automatically, not recorded
The 30-day returns rule	in the service passport, with a version and a date of effect	on the website, in the terms of sale, in the support handbook
Organisation AI Policy	in place, versions preserved	none
AI Service Passport	in place: time limit, grounds for refusal, criteria, sources, output map	none
AI Receipt for each interaction	in place, customer-facing part and machine-readable technical record	none
Service change register	in place	none
Dialogue log	part of the technical record of the receipt	on the platform, not linked to versions
Order management system, customer database	in place, operations linked to receipts	in place

Author of the analysis and of the architecture: Sergei Ponomarev, AI Business. The published methodology for test purchases of AI agents is freely available at aibusiness.vc/library. This document is a thought experiment. It is not a certification, an audit, a validation of any method, or an assessment of any real company.