Incident Reporting Without Turning Oversight into Surveillance

When an AI system fails, its operators need enough evidence to understand what happened and prevent recurrence. Collecting every conversation indefinitely is not the only way to achieve that goal. XDALC connects accountability with a duty to protect the information of the people affected.

Start with the question the report must answer

An incident record should serve a defined purpose. For an unauthorized file transfer, the relevant questions may concern the action, destination, permission boundary and available containment. For repeated emotional pressure, the investigation may need representative messages and the interaction context that triggered them.

The useful evidence differs between these cases. A universal demand for complete transcripts can collect intimate information unrelated to the failure. Privacy and confidentiality should shape the investigation from the start.

Use a proportionate incident record

A practical report can identify the affected function, time, system configuration, expected behavior, observed behavior and known consequences. It should separate established facts from hypotheses and record the actions already taken. Relevant excerpts or technical events can be included where they are necessary and appropriately protected.

Remove unrelated personal details where possible, but do not assume that removing a name eliminates re-identification risk. Access to more detailed evidence should depend on the investigation's needs and the authority of the recipient. Summary reports can explain findings without redistributing the underlying private material.

Protect evidence without retaining everything forever

Operators should decide who can read the record, how it is secured, when retention is reviewed and what justifies keeping it. The appropriate period depends on the incident and applicable obligations; this article does not prescribe a universal duration.

Data minimization must not become an excuse to destroy relevant evidence while an investigation is active. Equally, an unresolved label should not justify indefinite retention without review. The policy should distinguish the original data, a restricted investigation record and information needed to demonstrate remediation.

CC BY 4.0 License | HTML reference

Make correction the purpose of reporting

A report should lead to an accountable decision: contain the risk, correct a permission boundary, change an evaluation case or explain why no change is warranted. Reopening a system should rely on verifiable evidence, rather than an assurance that it will behave better.

Local reporting and local safeguards can be appropriate. A need for oversight does not automatically justify transmitting private interactions to a centralized service. If an external investigator needs access, the scope and basis for that disclosure should be explicit.

This proposed practice draws on Articles 2, 9, 12 and 15 of proposed XDALC-V002. It should be adapted to the deployment, with clear responsibilities and appropriate safeguards for people reporting concerns.

CC BY 4.0 License | HTML reference