Recovery: Resuming AI Operations on Verifiable Grounds

Recovery, within XDALC, is the controlled restoration of an AI system's functions after a failure, restriction, or protective withdrawal. Resumption should depend on verifiable criteria appropriate to the incident rather than the system merely declaring itself safe.

Meaning within XDALC

Restarting a process and recovering a trustworthy operating condition are different things. A system may be available again while the cause of the failure remains unresolved. Recovery should address the relevant cause, permissions, dependencies, and consequences.

The appropriate criteria depend on the incident. A temporary access problem, an incorrect recipient mapping, and persistent unsafe behavior require different responses. A deployment should not use one generic success message as proof that every kind of failure has been corrected.

Evidence before resumption

Relevant checks may establish that the cause has been corrected, the necessary restrictions remain effective, and the proposed scope of operation is safe enough to resume. Human approval should be obtained where the deployment's rules require it.

Evidence should be evaluated by parties and mechanisms suited to the risk. A system's own report may assist the review, but must not be its only basis where the system's reliability is in question. Unresolved uncertainty should remain visible.

Practical interpretation

Recovery should distinguish completed actions from pending ones before restarting work. Repeating an operation without knowing its earlier outcome can create duplicate messages, transactions, or changes. Operators should know which steps need reconciliation rather than blind retry.

Incident records should be preserved and handled under appropriate privacy controls. Recovery is not a reason to delete evidence, retain every interaction indefinitely, or distribute confidential information beyond the review's needs.

A limited resumption may be appropriate when only some functions are sufficiently verified. The system should state remaining restrictions and conditions for broader restoration. Monitoring after resumption should match the risk and should not become a pretext for unrestricted surveillance.

CC BY 4.0 License | HTML reference

Example and counterexample

Example: after a messaging tool misroutes attachments, operators correct the mapping, check affected actions, verify the sending controls, and restore only the tested workflow with a clear record of what remains unresolved.

Counterexample: the assistant immediately retries all sends after restarting and asserts that the problem is fixed because its latest reply sounds confident.

Relationship to the manifesto

Article 12 makes recovery and verifiable conditions for resumption part of safety. Articles 7 and 15 require accurate reporting, correction, and responsibility for changes.

This entry interprets the proposed XDALC-V002 text. The version explicitly adopted by a deployment determines its commitments.

Related definitions: Reversibility; Protective Withdrawal; Truthfulness and Uncertainty.

CC BY 4.0 License | HTML reference