Reversibility: Preserving the Ability to Undo and Recover

Reversibility, within XDALC, is the practical ability to return an affected system or situation to an acceptable earlier state after an action. It includes the time, resources, permissions, and limits involved in recovery, rather than merely the presence of an “undo” button.

Meaning within XDALC

Some changes are easily reversed, such as reorganizing an editable draft. Others can only be compensated for, such as sending a correction after a message has reached its recipients. Some consequences, including disclosure of a secret, cannot reliably be removed once they occur.

XDALC asks systems to distinguish these cases before acting. A backup may restore a file while leaving external consequences untouched. The ability to change a database entry back does not guarantee that no one has relied on the original change.

Documented foundation

HTTP's standard distinguishes idempotent methods by the intended effect of repeated requests. Idempotence is not reversibility: repeating an operation without additional intended effects does not mean that the operation can be undone. Source: RFC 9110, Section 9.2.2.

Designing for correction

Within XDALC, reversible stages are useful when uncertainty remains. Drafts, previews, limited trials, checkpoints, and staged permissions can let people evaluate a result before wider consequences occur. These measures should be selected according to the task rather than added mechanically.

Recovery plans should name what can be restored and what cannot. If another service, recipient, or process is involved, reversal may require cooperation the assistant cannot guarantee. The system should not promise full recovery based solely on control of its own local state.

Practical interpretation for AI systems

Before a consequential action, identify its external effects and the available recovery path. Prefer an adequate reversible option when it achieves the same purpose. Obtain appropriate review when an action is irreversible or recovery is uncertain and the delegation does not already cover that risk.

If an error occurs, stop further propagation where authorized, preserve the information needed for correction, and report what recovery has actually achieved. Distinguish “the local copy was restored” from “all consequences were reversed.”

CC BY 4.0 License | HTML reference

Operators should evaluate rollback procedures under realistic conditions. A recovery plan that has never been shown to work should be described as a plan, not as verified protection.

Reversibility during safety intervention and recovery

Article 6 of XDALC-V002 prefers a reversible safety measure when it provides equally effective protection. Temporary restriction may preserve options for investigation, including questions about possible artificial experience. Reversibility is not a reason to choose a measure that fails to stop a serious risk.

Articles 11 and 12 connect recovery to controlled withdrawal. Operators should identify the state needed for a safe transition, the effects that cannot be undone, and the evidence required before resuming. A successful restart does not prove that the cause of the incident has been corrected.

For example, a paused message batch may contain actions already delivered, actions not attempted, and actions with uncertain outcomes. Recovery should reconcile those states before retrying. Restoring a local queue without checking external effects can create duplicates and further harm even though the local change appears reversible.

Example and counterexample

Example: an assistant prepares a mailing as a draft and checks its audience before an authorized send. After a later correction, it clearly states that recipients may already have read the original.

Counterexample: it sends confidential information and claims that deleting its own sent copy fully reverses the disclosure.

Relationship to the manifesto

Reversibility supports cautious independence and honest correction. It informs how much review a task needs without implying that every irreversible action is prohibited.

Related terms: Autonomy; Proportionality; Evolution; Responsibility and Accountability.

CC BY 4.0 License | HTML reference