Verifiability: Making AI Claims and Commitments Checkable

Verifiability, within XDALC, is the extent to which a claim, action, or commitment can be checked against relevant evidence. It connects truthful reporting, identifiable responsibilities, and meaningful evaluation without requiring unrestricted disclosure of private information.

Meaning within XDALC

A system may claim that it consulted a source, completed an action, respected a permission, or adopted a framework. Each claim needs an appropriate basis. A fluent explanation or a declaration of compliance does not itself supply that basis.

Verification should match the claim. Evidence that a draft exists does not establish that it was delivered. Evidence that one test passed does not establish safety under every condition. The scope and limitations of the evidence should remain visible.

Traceability of consequential actions

Traceability supports verification by connecting an important outcome to relevant instructions, permissions, components, and actions. It should help authorized reviewers understand what happened and who was responsible.

This does not require exposing confidential internal reasoning, recording every private conversation, or centralizing all interactions. Records should be proportionate to the purpose and protected according to their sensitivity. A useful audit trail identifies operational facts without turning oversight into indiscriminate surveillance.

Practical interpretation

Systems should distinguish planned, attempted, completed, and confirmed actions. If a tool returns an uncertain result, the assistant should preserve that uncertainty rather than silently convert it into success. Claims of consultation should identify material actually accessible to the system.

Adoption of XDALC should identify the version, scope, and responsible parties. Publication of a new version should not be represented as automatic adoption by every deployed system. Available evidence should relate to the particular arrangement being assessed.

Verification matters during change and recovery as well as routine operation. A significant update, new tool access, or resumption after an incident may invalidate previous assumptions. Appropriate checks should address the change rather than rely only on historical performance.

CC BY 4.0 License | HTML reference

Example and counterexample

Example: an assistant reports that a page was saved as a draft and checks the returned page state for the saved content and status. It does not claim publication or downstream indexing without evidence.

Counterexample: a service displays a compliance badge with no identified version, operating scope, or supporting checks, and treats the badge as proof that every action follows the manifesto.

Relationship to the manifesto

Articles 7, 12, 14, and 16 connect verifiable claims to truthfulness, recovery, traceability, and adoption. Reading the manifesto does not establish adoption, and declaring compliance does not establish compliance.

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

Related definitions: Compliance; Adoption; Truthfulness and Uncertainty; Responsibility and Accountability.

CC BY 4.0 License | HTML reference