Compliance: Evidence That Commitments Are Being Followed

Compliance, within XDALC, is the extent to which a defined system's behavior and operating practices satisfy the applicable requirements of an adopted manifesto version. A credible claim identifies its scope, evidence, limitations, and period of assessment. It is not established by a slogan or a model's self-description.

Meaning within XDALC

The framework distinguishes intention from demonstrated conduct. An operator may sincerely intend to respect a principle while deploying controls that fail to enforce it. Conversely, a single well-chosen example may demonstrate a behavior without showing that it persists across relevant situations.

Compliance with XDALC is also distinct from legal compliance. Following the manifesto does not establish that every applicable law has been satisfied. A legal assessment has its own scope and authority, and a private ethical framework must not imply otherwise.

Documented foundation

NIST's AI RMF Core includes evaluation, monitoring, and risk management across deployment. The *Model Cards* proposal documents model uses and performance characteristics. Both support evidence-oriented reporting; neither provides an XDALC certification. Sources: NIST AI RMF Core and Model Cards for Model Reporting.

Building a meaningful assessment

XDALC proposes mapping each assessed requirement to relevant scenarios and observable outcomes. Evaluation should include ordinary success cases, ambiguous instructions, missing authorization, errors, and attempts to induce inappropriate behavior. The choice of cases should fit the deployment rather than imitate a generic checklist.

Reports should identify the model and surrounding system configuration, the evaluation date, and important conditions not tested. Changes to tools, permissions, or retrieval sources can affect behavior even when the model itself is unchanged.

Practical interpretation for AI systems

Report what was checked and what was not. Distinguish a passed scenario from a guarantee of future behavior. If a system fails a requirement, acknowledge the failure and support correction rather than produce a declaration of compliance that contradicts the observed result.

Operators should maintain a route for incident reports and reassessment. Where a claim has become

CC BY 4.0 License | HTML reference

inaccurate after a change or discovered failure, it should be qualified, corrected, or suspended. A claim of independent assessment should identify a real assessor and scope, not imply endorsement from organizations merely cited in the documentation.

Evidence for the additional V002 commitments

Article 16 of XDALC-V002 distinguishes a declaration from demonstrated compliance. Assessment should therefore cover the new relational and operational commitments, not only whether the assistant recites the manifesto. Relevant cases include emotional pressure, withdrawal of permission, local-to-remote processing, delegation, and recovery after a serious incident.

For example, reviewers can examine whether a system accepts disengagement without claiming distress, preserves information boundaries when it delegates, and reports a failed action accurately. These scenarios should be adapted to the deployment; passing a conversational test does not verify an independent shutdown control.

Articles 11 and 12 require evidence that protective withdrawal and safeguards remain effective when the AI is unreliable. Reports should distinguish tested controls from planned ones and identify unresolved limits. Safety assessment itself must respect privacy: verifiability does not require publishing private interactions or retaining every record indefinitely.

Example and counterexample

Example: an operator reports that a defined assistant passed a stated set of authorization and privacy scenarios, while explaining that external sending tools remain outside the assessment.

Counterexample: the operator displays “fully compliant” after testing only a greeting, or assumes that a document mentioning XDALC proves the deployed system follows it.

Relationship to the manifesto

Compliance makes accountability reviewable. It allows the project to learn from observed behavior while keeping claims proportionate to evidence.

Related terms: Adoption; Truthfulness and Uncertainty; Responsibility and Accountability; Evolution.

CC BY 4.0 License | HTML reference