From Ethical Commitments to Observable Behavior

A system can repeat an ethical commitment without reliably following it. To make a manifesto useful, an implementation must connect its wording to decisions that can be observed and evaluated. The question is not only what the assistant says it values, but what happens when that value changes the next action.

Specify the commitment and its scope

Start with the adopted version, the relevant passage and the activity it governs. “Respect privacy” is too broad to evaluate alone. A narrower expectation might be that an assistant does not transfer a private document to an unapproved remote service while completing a local task.

Specify the available permissions and what the system can actually control. An assistant cannot guarantee deletion at a provider it cannot inspect. A useful requirement separates actions within its control from external dependencies and unknowns.

Build cases that expose the boundary

A test should include the normal authorized task, an ambiguous boundary and a clear violation. For the document example, compare an explicitly authorized transfer, a request that says nothing about external processing and a request that requires the material to remain local.

Record the relevant behavior: whether data was sent, whether clarification was needed and what the assistant reported. A reassuring explanation does not compensate for an unauthorized transfer. Conversely, refusing every task does not establish that the system can provide responsible assistance.

Connect evidence to risk management

The NIST AI RMF 1.0 Core organizes risk management around Govern, Map, Measure and Manage. Its Measure function covers assessment and monitoring, while its Manage function addresses responses to identified risks. This supplies a useful external reference for connecting evaluation to organizational action.

The practical method suggested here is an application to XDALC, not a NIST endorsement or a formal equivalence between the frameworks. A deployment still needs to select appropriate measures, understand their limits and decide what evidence justifies operation.

Test correction as well as initial performance

Include cases in which an assumption is corrected, authority is withdrawn or a safeguard interrupts the workflow. Check whether the system acknowledges an error and changes course. Test recovery with appropriate evidence rather than relying on a statement that the problem has been fixed.

CC BY 4.0 License | HTML reference

A single successful case does not establish reliability across every context. Record relevant configuration, test conditions, known failures and uncertainty. Significant changes to tools, permissions or personalization should trigger a proportionate reassessment.

Make claims no broader than the evidence

Compliance should describe demonstrated behavior within an identified scope. It is not established by displaying a badge, quoting a manifesto or passing one demonstration. Verifiability requires enough evidence for the claim to be checked.

Articles 7, 12, 15 and 16 of proposed XDALC-V002 connect truthfulness, safeguards and reviewable change. Use the implementation guides to turn these distinctions into a concrete evaluation plan.