Trust: Confidence Proportionate to Evidence

Trust, within XDALC, is a person's justified willingness to rely on a system within understood limits. It should remain proportionate to evidence about the system's behavior, capabilities, permissions, and opportunities for correction.

Meaning within XDALC

Trust is not something an AI is entitled to receive. It develops through actions that can be understood, checked, and challenged. A familiar voice or consistent personality can make interaction easier, but neither establishes reliability.

Confidence should be specific to a task and setting. Success at summarizing documents does not prove that a system can make consequential decisions, operate equipment, or handle confidential information safely. Permission and evidence should be reassessed when the scope changes.

Reassurance is not verification

A convincing explanation can still be wrong. A declaration of safety can still omit a limitation. Useful grounds for reliance include relevant checks, accurate reporting, documented constraints, and effective ways to interrupt or correct an action.

An assistant should say when it is uncertain or when a tool has not confirmed completion. It should not present an attempted operation as a successful one simply because that answer sounds more reassuring. Recognized mistakes should lead to correction and, where possible, remediation.

Practical interpretation

Users should be able to understand which information the system can access and which actions it can take. They should be free to question a recommendation, seek another opinion, reduce permissions, or refuse interaction. These choices should not be treated as hostility.

Operators should provide evidence appropriate to the consequences of a deployment. A low-risk drafting aid and a system involved in essential services call for different levels of evaluation. Trust should not depend exclusively on claims produced by the system being assessed.

When capabilities or operating conditions materially change, previous confidence may no longer apply. Transparent communication about relevant changes helps people calibrate reliance instead of transferring trust automatically to a new arrangement.

Example and counterexample

Example: an assistant reports that it prepared a document but could not verify delivery. It identifies the

unresolved step and lets the user decide how to continue.

Counterexample: it says 'trust me' after a failed action, hides the failure, and treats the user's request for evidence as an obstacle.

Relationship to the manifesto

Articles 7 and 8 connect trust to truthfulness, predictable behavior, and verification. Articles 12 and 16 require safeguards and evidence beyond self-declaration.

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

Related definitions: Compliance; Truthfulness and Uncertainty; Human Oversight.