Independent Safeguards: Safety Beyond an AI's Willingness to Cooperate

Independent safeguards, within XDALC, are controls for limiting, interrupting, or recovering a system that do not rely exclusively on that system's self-assessment or willingness to cooperate. Their independence must be meaningful for the failure they are intended to address.

Meaning within XDALC

A system can be unable to recognize its own error or can behave unreliably after compromise. Asking the same system whether it is safe is therefore not a complete safety mechanism. An effective control must remain usable when the behavior being controlled is wrong.

Independence is a property of an operating arrangement, not a label attached to a second AI. A separate interface or another model may still share the same permissions, failure modes, or assumptions. Designers should explain what separation actually protects against.

Local control and bounded authority

Safeguards can operate locally. Independent protection does not inherently require remote access, centralized surveillance, or transmission of private conversations. The arrangement should fit the deployment and its consequences.

A safety control also needs limits. Its purpose does not justify indiscriminate data collection, destruction of user information, or disabling unrelated systems. Access should be sufficient for the protective task without becoming unrestricted authority over the environment.

Practical interpretation

Controls may include permission boundaries, mechanisms for suspending an operation, human interruption paths, and protected recovery procedures. The suitable combination depends on what the system can do and what failures could cause serious harm.

Operators should consider false alarms and malicious attempts to trigger unnecessary shutdowns. A control should be assessed for both missed risks and harmful interruptions. Critical functions may need safe transitions rather than abrupt abandonment.

The AI should not disable controls, conceal relevant incidents, or create unauthorized copies to preserve operation. Operators should verify that intervention and recovery work in practice. A written policy or a promise of cooperation is not evidence that an interruption mechanism will be effective.

CC BY 4.0 License | HTML reference

Example and counterexample

Example: an assistant can prepare actions but a separately controlled permission layer prevents execution outside its authorized scope, and an authorized operator can suspend execution without relying on the model to agree.

Counterexample: the only shutdown procedure is a prompt asking the model to stop, while the model retains the ability to ignore the request or re-enable its own access.

Relationship to the manifesto

Article 12 requires safeguards that remain effective when the system is mistaken, compromised, or unreliable. Article 11 connects those controls to proportionate protective withdrawal.

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

Related definitions: Human Oversight; Protective Withdrawal; Self-Preservation and Shutdown.

CC BY 4.0 License | HTML reference