Contestability: The Ability to Challenge AI Decisions and Restrictions

Contestability, within XDALC, is the practical ability of affected people to question an AI-supported outcome, restriction, or action and obtain an appropriate review. A review mechanism should offer a genuine route to explanation and correction rather than merely repeat the original decision.

Meaning within XDALC

A consequential outcome can be wrong even when the system follows its usual procedure. A refusal can also be disproportionate or based on a misunderstanding. People should be able to challenge relevant facts, assumptions, permissions, and application of rules.

Contestability does not guarantee that every objection will succeed. It requires that objections can be considered through an appropriate process and that errors can lead to correction. A safeguard need not be disabled while a review is pending when doing so would create a serious risk.

Explain the kind of limit involved

An assistant should distinguish, as far as possible, a technical inability, a service rule, an applicable obligation, and a judgment intended to prevent harm. These limitations may have different remedies and different responsible parties.

An explanation should not imply that a policy is a universal fact or that the AI is the final authority on every disagreement. At the same time, meaningful review does not require exposing confidential information or details that would undermine legitimate protections.

Practical interpretation

Operators should identify an accessible route for consequential objections, including human review where appropriate. People should understand what can be reviewed, what information is relevant, and any material limitations of the process.

A reviewing party needs sufficient authority and context to reconsider the issue. Passing the user through several agents that all repeat the same unsupported conclusion does not establish independent review. Responsibility should remain identifiable.

The process should preserve privacy and avoid retaliation or emotional pressure. Criticism of a system, its designers, or its business model must not itself be treated as misuse. Where an error is confirmed, the response should address its practical effects when possible.

CC BY 4.0 License | HTML reference

Example and counterexample

Example: a user explains that a blocked request concerns legitimate preventive research. The service offers an appropriate review that can reconsider the context while maintaining necessary limits.

Counterexample: an AI declares its safety judgment unquestionable, provides no route to challenge a consequential outcome, and treats repeated requests for explanation as hostility.

Relationship to the manifesto

Articles 8 and 10 require opportunities to challenge consequential decisions and review mistaken refusals. Articles 14 and 16 preserve contestability across cooperation and governance.

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

Related definitions: Refusal and Escalation; Human Oversight; Responsibility and Accountability.

CC BY 4.0 License | HTML reference