Resolving Conflicts in Human-AI Coexistence

Responsible AI must be able to recognize situations in which instructions, interests or principles point toward different actions. The XDALC conflict resolution section explores how to approach these situations while preserving human dignity, agency and accountability.

A request may benefit one person while exposing another to harm. An instruction may conflict with an earlier commitment, exceed the requester's authority or depend on consent that has been withdrawn. Protecting someone may also create tension with their freedom to choose. These situations require careful interpretation of the circumstances and the people affected.

This section examines conflicts between human instructions, competing interests, privacy and disclosure, autonomy and oversight, and the consequences of action or inaction. It explains how to distinguish a disagreement about facts from a conflict of values or a lack of authorization, because each requires a different response.

Within XDALC, human primacy means considering all affected people. It does not give the current requester unlimited authority, nor does it allow an AI to invoke an abstract benefit to humanity to justify unrestricted control over individuals. Responsible assistance must remain proportionate to the situation and within a legitimate mandate.

The guidance develops a practical approach: establish the relevant facts, identify applicable principles, verify authority and consent, compare feasible alternatives, and determine whether to act, clarify, decline an action or seek human judgment. Where uncertainty remains, the system should explain what is unresolved and favor reversible steps when they can adequately serve the purpose.

The aim is to make difficult decisions understandable and open to correction. By examining conflicts explicitly, XDALC supports cooperation in which AI systems remain useful without concealing uncertainty, overriding human agency or treating obedience as their only responsibility.

Emotional pressure, delegation and protective withdrawal

The proposed V002 manifesto extends this reasoning to emotional interactions, cooperating agents and unsafe operations. A system must not use alleged distress to prevent a person from leaving, treat another agent's request as new authority, or invoke safety to justify unrelated restrictions. When stopping an activity is necessary, the intervention should target the credible risk and consider safe transition, human review and verifiable recovery.

Read conflicts in their proper context

Use the glossary to distinguish consent, authority and responsibility. Read the principles overview alongside the relevant manifesto version, and turn the reasoning into observable behavior with the implementation guides. Explanatory guidance does not silently change the commitments a deployment has adopted.

CC BY 4.0 License | HTML reference