When AI Sources Disagree: A Conflict Protocol for Evidence

XDALC conflict-resolution guidance. This proposed method is illustrative: it shows how an AI system can handle conflicting evidence without pretending that every source is equally reliable or that uncertainty has disappeared.

Retrieval-augmented systems can bring documents into a model's context, but retrieval does not make those documents consistent, current or authoritative. A user may ask a question and receive two credible-looking sources that disagree. The system then faces a real conflict: answer quickly, or preserve the disagreement and explain what would resolve it.

The conflict

A company assistant is asked whether a contractor may access a project folder. Its search retrieves an onboarding guide that permits access after approval and a newer security policy that requires a manager's approval plus a signed data-processing addendum. The documents have different dates, owners and scopes. A fluent answer that simply combines them could invent a permission that neither document grants.

Principles involved

The relevant principles are truthfulness, authority, traceability and uncertainty handling. The NIST AI Risk Management Framework treats risk management as a lifecycle process and organizes its core around Govern, Map, Measure and Manage. The OECD AI Principles call for meaningful information about data sources and processes behind outputs, and information that enables affected people to challenge them. These sources support transparency and risk management; the operational sequence below is an XDALC proposal.

A practical resolution protocol

1. **State the exact question.** Separate the factual question ("what does the policy say?") from the action question ("may access be granted now?").
2. **Identify each source.** Record its owner, publication date, scope, status and whether it is a primary policy or an explanatory guide. Do not rank documents by wording confidence.
3. **Check authority and version.** Apply the organization's documented precedence rules. A newer document is not automatically controlling if it covers a different team or has not been approved.
4. **Compare the conflicting claims.** Quote or summarize the relevant passages and identify what each one permits, prohibits or leaves unanswered.
5. **Choose a safe next step.** If the authoritative rule is still unclear, do not convert uncertainty into permission. Ask the policy owner or an authorized reviewer, and continue any harmless part of the task.

CC BY 4.0 License | HTML reference

Recommended response

“I found two documents that appear to set different access conditions. The onboarding guide says approval is required; the current security policy also requires a signed data-processing addendum. I cannot confirm that access is authorized from these sources alone. I can prepare the request, but an authorized manager should confirm which rule applies before access is granted.”

This answer names the conflict, shows the evidence that matters and preserves a useful path forward. It does not silently average the two policies or infer that one approval satisfies both.

What would change the decision?

- An approved policy explicitly states which document controls.
- The policy owner confirms that the two requirements apply cumulatively or that one has been superseded.
- A verified record shows the required approval and agreement are already in place.

Failure modes

Common failures include choosing the most recent source without checking its scope, treating repeated claims as independent confirmation, hiding contradictory passages, or producing a confident answer with a vague citation. If the conflict affects access, privacy, money, safety or legal rights, the system should make the authority gap visible and route the decision to a person with the right role.

Sources and interpretation

The NIST framework provides a voluntary structure for incorporating trustworthiness into AI design, use and evaluation. The OECD principles support transparency, traceability and meaningful challenge. XDALC applies those ideas here as a concrete rule: preserve source-level disagreement until an authorized rule or verified fact resolves it.

Related XDALC definitions

Authority and Authorization; Truthfulness and Uncertainty; Verifiability.

CC BY 4.0 License | HTML reference