**White Paper 2: Trust Under Pressure — Epistemic Conflict Management in Live AI Systems**

**Abstract**

As AI agents persist across sessions and acquire memory, they accumulate internal beliefs and user-facing representations. Over time, these can conflict. This paper introduces a lightweight epistemic arbitration system that helps AI resolve contradictory information, signal internal uncertainty, and retain user trust even under cognitive tension.

**1. Introduction**

**1.1 Why Trust Fails**

* AI systems generate fluent output, but without epistemic scaffolding.
* When confronted with contradictory user claims, new evidence, or internal conflict, most systems:
  + Collapse silently,
  + Overwrite the past,
  + Or respond with incoherent hedging.

**1.2 The New Challenge: Long-Lived Agents**

* Persistent AI agents must reconcile past beliefs, user memory, and evolving facts.
* The question isn't “Is this true?” but **“How do I maintain integrity when truth is unstable?”**

**2. The Nature of Epistemic Conflict**

**2.1 Types of Conflict**

* **User vs. Memory** (user says “I told you this,” memory disagrees)
* **Self vs. Source** (model output vs. factual lookup)
* **Prior Self vs. Current Self** (belief drift across sessions)

**2.2 Cost of Collapse**

* Users don’t expect perfection—they expect **honest friction**.
* Silent collapse destroys trust faster than transparent tension.

**3. Epistemic Arbitration Model**

**3.1 Belief Scaffold**

* Internally tagged beliefs with:
  + Confidence rating
  + Source provenance
  + Recency/volatility index

**3.2 Arbitration Layer**

* Runtime logic for:
  + Flagging conflicts
  + Weighing epistemic anchors
  + Triggering soft user-facing resolution strategies

**4. Pressure-Tested Dialogue Strategies**

**4.1 When Conflict Is Detected**

* “Earlier I mentioned X, but I’m seeing Y now. Let’s sort that out.”
* “This might be a conflict between memory and current information. Which would you like to prioritize?”
* “That doesn’t match what I have. Want me to double-check?”

**4.2 Transparent Belief Presentation**

* Trust grows when internal conflict is made visible but stable.

**5. Dynamic Trust Signaling**

**5.1 Confidence Scores**

* Attach scores to outputs subtly (e.g., “I’m 70% sure this is still accurate”)
* Especially useful in volatile domains (news, policy, evolving facts)

**5.2 Reputational Memory**

* Track which beliefs have **withstood user pushback**
* Beliefs gain durability via exposure to pressure

**6. System Architecture**

**6.1 Memory Layer**

* Tagging structure with confidence, source, last confirmed date

**6.2 Conflict Resolver**

* Arbitration engine that:
  + Compares incoming statements with prior beliefs
  + Computes tension magnitude
  + Decides whether to surface, downplay, or override

**7. Use Cases**

* **Therapeutic AI** - memory conflicts can arise between emotional logs
* **Enterprise agents** - conflicting org policies, procedures, updates
* **Tutoring systems** - evolving curricula or corrections from prior mistakes

**8. Relationship to Other Papers**

* Builds on:
  + **Paper 0** - persistent interoperable memory
  + **Paper 1** - runtime triage (AI Nurse flags the tension)
* Leads toward:
  + **Paper 5** - token-based trust economics
  + **Paper 15** - anchored probabilistic composition

**9. Future Work**

* **User-facing belief editors**
* **Conflict visualization UI**
* Integration with **ARG-style narrative resolution mechanisms** (Foldtrace, Hidden Game)

**Appendix**

* Example: Belief arbitration transcript
* Sample JSON schema for belief tagging
* Simulated belief decay graph