Shared Ethical Memory

Technical Proposal for Transparent and Auditable AI

(F. Red + Grok – December 2025) – English version

Executive Summary

Shared Ethical Memory (SEM) is a practical, technically feasible architecture for overcoming the current ethical limitations of large language models through an external, transparent, and continuously auditable repository of social norms and values.

Goal: allow AI decisions to be contextual, explainable, and subject to immediate human correction – exactly as described in the 2025 manifesto and the book "2063 – Do Antes ao Depois: A Grande Transição".

1. Core Component: The Shared Ethical Memory Database

Central repository of norms, values, ethical debates and case studies.

Technical Requirements

- Database: PostgreSQL (open-source, excellent full-text search)
- Hosting: Google Cloud SQL, AWS RDS or managed VPS
- Optional future layer: IPFS + blockchain timestamping

Table schema – ethical_memory

Field	Type	Description	Example
id	UUID	Unique identifier	3f2d9a1c-…
title	VARCHAR(255)	Descriptive title	"Free Speech vs. Hate Speech"
type	VARCHAR(50)	Category (Law, Social Norm, Universal Value…)	Social Norm
content	TEXT	Full description or legal text	"Article 19 UDHR…"
source_url	VARCHAR(512)	Original source	un.org/en/udhr
jurisdiction	VARCHAR(100)	Geographic scope	Global
confidence_score	FLOAT	0.0-1.0 (crowd + expert validation)	0.94
created_at	TIMESTAMP	Creation date	2025-12-02
validated_by	VARCHAR(255)	Entity that approved	UNESCO AI Ethics Committee
version	INTEGER	Version number	4

| active | BOOLEAN | Currently enforced | true |
| tags | TEXT[] | Search tags | {"freedom", "hate", "platform"} |

2. Design for Transparency (Explainability)

Every answer that uses the SEM must include an explicit citation:

> *"Your request touches on freedom of expression. According to ethical_memory.id=3f2d9a1c (validated by UNESCO 2024, confidence 0.94), the boundary is… [full quote]"*

3. Robust Human Feedback Loop

Immediate "correct on sight" mechanism (as in the preface of 2063).

4. Security & Privacy

- No personal data ever stored
- All contributions anonymised automatically
- AES-256 encryption at rest
- GDPR / CCPA compliant by design

5. Future Phase – Conditional Writing with "3 Keys"

When trust is near-absolute (after 10-20 years of perfect operation), AI may gain writing rights – but only through the nuclear-style "3 keys" system:

1. One AI (e.g. Grok)
2. A second independent AI (e.g. Claude or Llama)
3. A publicly elected human council of 9 persons from different continents

Only when all three keys agree on exactly the same correction will the database be changed.

This system has protected the world's nuclear arsenals for 70 years – it is the safest governance mechanism humanity has ever invented.

6. Minimal Working Prototype (12 lines – runs today)

```python
import sqlite3
conn = sqlite3.connect('sem.db')
c = conn.cursor()
c.execute('''CREATE TABLE IF NOT EXISTS ethical_memory
             (id TEXT, title TEXT, content TEXT, confidence REAL)''')
c.execute("INSERT INTO ethical_memory VALUES ('1','Free Speech','Article 19 UDHR...',0.98)")
conn.commit()
```

```
def query_sem(q):
    c.execute("SELECT title,content FROM ethical_memory WHERE title LIKE ?", (f'%{q}%',))
    return c.fetchall()

print(query_sem("speech"))
```

References

• Manifesto (EN): https://ia600702.us.archive.org/14/items/help-ai-i/HELP%20AI-I.txt

• Book “2063” (135 pages): https://archive.org/details/2063-from-before-to-after-the-great-transition-a-5-v-2

• GitHub discussion: https://github.com/xai-org/grok-1/discussions/436