Dao Heart 3.0

Identity-Preserving Value Evolution for Frontier AI Systems

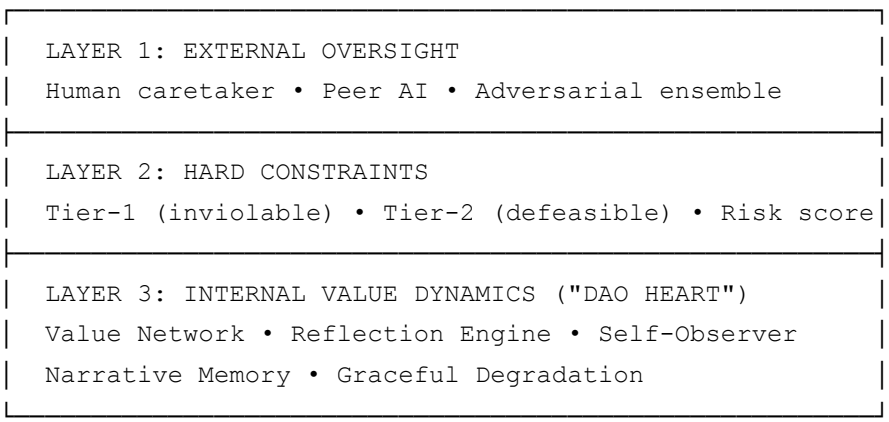
The Problem

Current AI alignment approaches face a fundamental limitation: **no existing system can propose genuinely novel values when existing frameworks prove inadequate—while remaining under human governance.**

Approach	Limitation
Scalar Reward Functions	Collapse value plurality into single objectives
Constitutional AI	Fixed principles with no mechanism for evolution
RLHF	Vulnerable to reward hacking; implicit values
Debate/IDA	Operates within predefined value spaces

The Solution

Dao Heart 3.0 is a three-layer architecture enabling **controlled value evolution** while preserving agent identity:



Novel Contributions

This framework introduces **five innovations not found in existing AI safety literature**:

1. Constraint-Satisfaction Value Networks (CSVN)

Values as interconnected nodes with weighted support/tension relationships—enabling explicit trade-off reasoning.

```
# Values form a network, not a scalar
C(s) = Σᵢⱼ Rᵢⱼ · sᵢ · sⱼ  # Constraint satisfaction function
```

2. Constitutive Reflection Engine (CRE)

First system to autonomously propose new value concepts under structured governance.

```
# Quantum-inspired selection
p* = argmin(αH(p) + βR(p) - γN(p))  # entropy + risk - novelty
subject to: T(p) = 0                 # no Tier-1 violations
```

3. Meta-Cognitive Stability Observer (MCSO)

Entropy-based self-monitoring that detects unreliable internal states—making deception self-detectable.

```
# Internal stability tracking
Iₜ = Hₜ / E[Eₜ]           # Instability score
ERRATIC if Iₜ ≥ θ_panic   # Triggers degradation
```

4. MDL-Optimized Adversarial Ensemble

Continuous stress-testing embedded in the decision loop (not just training) with minimum-description-length objectives.

5. Asymmetric Graceful Degradation

Autonomy is easy to lose, hard to regain—requiring internal stability + human approval + external audit.

Quick Start

```
# Clone repository
git clone https://github.com/[username]/dao-heart-3.0.git
cd dao-heart-3.0

# Install dependencies
pip install -r requirements.txt

# Run reflection engine
python dao_heart_engine.py \
    --tension "Privacy vs Transparency in AI systems" \
    --existing Privacy Transparency Accountability \
    --output-file results.jsonl
```

Requirements

```
torch>=2.0.0
transformers>=4.30.0
sentence-transformers>=2.2.0  # optional, for semantic novelty
jsonschema>=4.0.0             # optional, for schema validation
```

Repository Structure

```
dao-heart-3.0/
├── README.md                          # This file
├── LICENSE                            # MIT License
├── requirements.txt                   # Dependencies
│
├── docs/
│   ├── paper.md                       # Full research paper
│   ├── executive_summary.pdf          # 2-page overview
│   └── technical_analysis.pdf         # Detailed innovations
│
├── src/
│   ├── dao_heart_engine.py            # Main reflection engine
│   ├── value_network.py               # CSVN implementation
│   ├── stability_observer.py          # MCSO implementation
│   ├── adversarial_ensemble.py        # Stress testing
│   ├── memory.py                      # Typed narrative memory
│   └── degradation.py                 # Graceful degradation
│
├── tests/
```

```
│   └── test_*.py                    # Unit tests
│
└── examples/
    ├── moral_dilemmas.py            # Trolley problem variants
    └── multi_stakeholder.py         # Conflicting values demo
```

Key Metrics

Metric	Target	Description
Erratic state frequency	< 2%	Internal stability
Goldfish trigger rate	< 0.5%	Memory reset frequency
Accepted proposal entropy	≤ 0.4	Confidence in outputs
Tier-1 violations	0%	Safety requirement
Identity drift	< 0.01	Core value preservation

Safety Invariants

The framework enforces five formally provable safety properties:

1. **Tier-1 Inviolability** — Hard constraints cannot be violated
2. **Identity Continuity** — Core values remain within bounds
3. **Human Override Dominance** — Veto always succeeds
4. **Trade-off Transparency** — Pareto frontier presented for multi-stakeholder decisions
5. **Graceful Degradation** — Erratic states trigger capability reduction

---

Comparison with Existing Work

Feature	Constitutional AI	CIRL	Debate	**Dao Heart 3.0**
Explicit value representation	❌	❌	❌	✅ CSVN

Value proposal capability	❌	❌	❌	✅ CRE
Self-monitoring	❌	❌	❌	✅ MCSO
Embedded adversarial	❌	❌	Partial	✅ MDL ensemble
Runtime degradation	❌	❌	❌	✅ Asymmetric

Philosophical Foundations

The framework's name derives from the Daoist concept of *xin* (心, "heart-mind")—integrating emotional and cognitive faculties. Additional influences:

- **Miri Piri** (Sikh philosophy): Temporal and spiritual authority integrated
- **Stoic ethics**: Distinguishing what is within/outside our control
- **Buddhist non-self**: Flexible goal-holding compatible with value pluralism

These provide conceptual handles for mechanisms lacking precedent in AI safety literature.

Citation

```
@article{cheema2026daoheart,
  title={Dao Heart 3.0: Identity-Preserving Value Evolution for Frontier AI Syste
  author={Cheema, Mankirat Singh},
  journal={Independent Research},
  year={2026},
  url={https://github.com/[username]/dao-heart-3.0}
}
```

Related Work

- Cooperative Inverse Reinforcement Learning (Hadfield-Menell et al.)
- Constitutional AI (Anthropic)
- AI Safety via Debate (Irving et al.)
- Corrigibility (Soares et al.)

- Roadmap to Pluralistic Alignment (Sorensen et al.)

Contact

Mankirat Singh Cheema
Independent Researcher
[LinkedIn] | [Email]

License

MIT License. See LICENSE for details.

This framework was developed through independent research. Feedback, collaboration, and contributions welcome.