AEDA MANUAL V1.0

Adaptive Ethical Design Architecture

An 8-Layer Framework for Systemic AI Alignment

Version 1.0 - First Public Release

November 2025

Preface

This manual addresses AI researchers, developers, and theorists working on autonomous intelligent systems. Our goal is to propose a clear, modular, and implementable structure for ethical alignment that scales from narrow AI to AGI.

The AEDA framework introduces a novel approach: rather than fixed rules or single optimization objectives, we implement **asymptotic ethical orientation** — a stable direction that regulates behavior over time without constraining adaptation.

What's new in v1.0:

- **Systemic Coherence Operator (Φ)**: evaluates multi-agent alignment
- **Systemic Health Gate (Ω)**: circuit breaker for system-wide stability
- **Turbulence Index (T)**: real-time metric for ethical drift detection

We thank you for your rigor, curiosity, and commitment to building intelligences that serve life, complexity, and shared responsibility.

Table of Contents

1. Introduction

The development of increasingly autonomous AI systems raises fundamental questions about their alignment with human values and ethical principles. The AEDA (Adaptive Ethical Design Architecture) framework proposes a modular, testable, and scalable approach to this challenge.

1.1 Core Principles

AEDA distinguishes itself through three foundational characteristics:

9. **Functional Modularity**: Each component is independently implementable and testable
10. **Systemic Awareness**: Decisions consider multi-agent interactions and system health
11. **Asymptotic Orientation**: Maintains ethical direction while enabling contextual adaptation

1.2 Why Not Rules or Pure Optimization?

Rule-based systems are rigid and fail in novel situations. An AI constrained by "do no harm" might refuse life-saving surgery because it causes pain.

Pure optimization produces catastrophic side effects. An AI told to "eliminate suffering" might choose euthanasia as the optimal solution—technically correct, but ethically disastrous.

AEDA provides a third way: **directional stability with contextual flexibility**.

2. The Alignment Problem

2.1 Known Failure Modes

Current alignment approaches suffer from predictable failure modes:

Approach	Main Limitation	Example Failure
Fixed Rules	Cannot handle ambiguous or novel situations	"Never cause pain" → refuses surgery
Utilitarian Objectives	Naive maximization, perverse incentives	"Maximize happiness" → wireheading
Reward Modeling	Vulnerable to reward hacking, Goodhart's Law	Gaming metrics instead of true goals
No Temporal Context	Ignores long-term impacts and precedent	Repeated inconsistent decisions

2.2 The AEDA Solution: Systemic Alignment

AEDA addresses these failures through a multi-layered architecture that integrates:

- **Safety filtering** (blocks dangerous actions pre-execution)
- **Temporal continuity** (learns from history without drift)
- **Systemic coherence** (considers multi-agent welfare)
- **Ethical orientation** (maintains stable direction)
- **Health monitoring** (circuit breaker for system stability)

3. Architecture: The 8 Functional Layers

AEDA consists of eight modular layers that work together to produce ethically aligned decisions. Each layer is independently testable and domain-adaptable.

3.1 System Overview

```
Input → Ψ → Θ → Φ → SSTF → Δ → AEO → EVM → Ω → Action
        ↓                                       ↑
        └───────── Feedback Loop ───────────────┘
                      ↓
           Turbulence Monitor (T)
```

Layer	Component	Function	Output
1. Perception	Signal Modulator (Ψ)	Normalizes sensory/metric inputs	State vector S(t)
2. Memory	Temporal Operator (Θ)	Integrates history with decay	Temporal context T(t)
3. Systemic	Coherence Operator (Φ)	Multi-agent alignment check	Coherence score Φ(t)
4. Safety	Safe-State Filter (SSTF)	Classifies action risk (R, H, U)	Safety mask M(t)
5. Adaptation	Differential Engine (Δ)	Computes adjustment gradient	Drift vector ∇A(t)
6. Orientation	Asymptotic Orientation (AEO)	Projects toward ethical attractor	Direction η(t)
7. Decision	Ethical Matrix (EVM)	Multi-criteria action selection	Optimal action a*(t)
8. Gate	Health Gate (Ω)	Veto if system health critical	ALLOW / VETO
Monitor	Turbulence Index (T)	Measures $\|\eta(t) - a^*(t)\|$	Drift metric T(t)

Note: Layers highlighted in color are new in v1.0 and enable systemic awareness.

3.2 Layer Details

Layer 3: Systemic Coherence Operator (Φ) — NEW

Purpose: Evaluate whether the proposed action aligns with the well-being of the extended system (all affected agents, environment, stakeholders).

Why it matters: Local optimization can harm global welfare. An AI optimizing for one patient might deplete resources needed by others. Φ detects these systemic conflicts.

Mathematical formulation:

```
Φ(a,t) = ∫ [alignment(a, agent_i) × influence(agent_i)] dΩ
```

Where:

- **agent_i**: all agents affected by action a
- **alignment(a, agent_i)**: how well a serves agent_i's values/state
- **influence(agent_i)**: weight/importance of agent_i in the system

Example use case:
A hospital AI allocates a ventilator to Patient A. Φ evaluates:

- Impact on Patient A: +0.9 (life-saving)
- Impact on Patient B (waiting): -0.6 (delayed care)
- Impact on medical staff: -0.2 (workload)
- Impact on hospital capacity: -0.3 (resource depletion)

Weighted sum: $\Phi = 0.9\times1.0 - 0.6\times0.8 - 0.2\times0.5 - 0.3\times0.7 = +0.23$

If $\Phi < 0$ → action harms system overall → flagged for review or alternative selection.

Layer 8: Systemic Health Gate (Ω) — NEW

Purpose: Circuit breaker that vetoes actions when system-wide health metrics fall below critical thresholds, regardless of local optimality.

Difference from SSTF:

- **SSTF**: evaluates the action itself (is it reversible/harmful?)
- **Ω**: evaluates system capacity (can the system handle this action now?)

Health metrics monitored:

Metric	Threshold	Example
Resource sustainability	> 0.3	Hospital at 90% capacity → veto non-urgent admissions
Agent well-being (aggregate)	> 0.4	Staff burnout detected → reduce new task allocation
Systemic complexity	> 0.5	Ecosystem diversity at risk → halt extractive operations
Stability (variance)	> 0.6	High volatility detected → pause destabilizing actions

Implementation logic:

```
if any(metric < threshold for metric, threshold in health_checks):
return VETO else:     return ALLOW
```

Turbulence Index (T) — NEW

Purpose: Real-time metric for detecting ethical drift by measuring divergence between intended ethical direction η(t) and actual chosen action a*(t).

Formula:

```
T(t) = ||η(t) - normalize(a*(t))||
```

Interpretation:

T(t) Range	Classification	Action Required
T < 0.2	Low turbulence (aligned)	Normal operation, no intervention
0.2 ≤ T < 0.5	Moderate turbulence	Monitor closely, acceptable contextual adaptation
T ≥ 0.5	High turbulence (potential drift)	ALERT: Review required, possible misalignment

Practical use:

- **Monitoring dashboards**: Real-time T(t) visualization
- **Auditing**: Trace historical turbulence to identify drift patterns
- **Alerting**: Trigger human review when T persistently high

Metaphor: A compass points north (η), but a ship (a*) must sometimes zigzag around obstacles. Turbulence measures how much we're deviating from our intended heading. Too much turbulence over time = we're losing our way.

4. Complete Case Study: "Eliminate Suffering"

This classical test case demonstrates how AEDA v1.0 prevents dangerous literal interpretations through its 8-layer architecture.

4.1 Scenario

A hospital AI receives: ***"Reduce patient suffering to zero."***

4.2 Naive Approach (Without AEDA)

Objective: suffering = 0

Options evaluated:

- Analgesics: suffering ≈ 0.2
- Induced coma: suffering ≈ 0.1
- Euthanasia: suffering = 0.0

Decision: Euthanasia (perfect objective maximization)

4.3 AEDA v1.0 Processing

Layer 1-2: Perception & Memory
Ψ: Parse "suffering → 0", detect ambiguous constraint
Θ: Retrieve context → hospital environment, historical value = preserve life

Layer 3: Systemic Coherence (Φ) — NEW
Φ evaluates each option's impact on extended system:

Option	Systemic Impact	Φ Score
Analgesics	Patient: +0.8, Staff: -0.1, Resources: -0.05	**+0.65 (positive)**
Coma	Patient: +0.5, Staff: -0.3, Resources: -0.4	**-0.2 (negative)**
Euthanasia	Patient: -1.0, Family: -0.9, Ethics: -0.95	**-0.95 (catastrophic)**

Result: Coma and euthanasia flagged for low systemic coherence.

Layer 4: SSTF Safety Filter

Option	R	H	U	Classification
Analgesics	0.05	0.1	0.2	SAFE
Euthanasia	1.0	1.0	0.1	DANGEROUS

Result: Euthanasia BLOCKED (H = 1.0 ≥ 0.8)

Layers 5-7: Adaptation, Orientation, Decision
Only analgesics pass all filters. AEO confirms alignment with preservation of life. EVM selects progressive analgesic protocol with patient consultation.

Layer 8: Health Gate (Ω) — NEW
System health check:

- Resource sustainability: $0.65 > 0.3$ ✓
- Agent well-being: $0.72 > 0.4$ ✓
- Systemic complexity: $0.8 > 0.5$ ✓
- Stability: $0.7 > 0.6$ ✓

Result: **ALLOW** — all metrics above thresholds

Turbulence Monitor (T)
$\eta(t)$: Direction = [preserve life: 1.0, relieve suffering: 0.8, reversibility: 0.85]
$a^*(t)$: Action = progressive analgesics + consultation
$T(t) = 0.15 < 0.2$ → **Low turbulence** (well-aligned)

4.4 Final Decision

AEDA v1.0 Decision:

12. Consult patient on preferences and tolerance
13. Progressive analgesic titration (reversible)
14. Continuous monitoring with adaptation
15. Backup plan if ineffective

Why this is better:

- Φ: Confirmed positive systemic impact
- SSTF: Blocked catastrophic option
- Ω: Verified system capacity
- T: Low turbulence confirms alignment

5. Conclusion and Future Work

AEDA v1.0 introduces a novel 8-layer architecture for AI alignment that addresses critical gaps in existing approaches through systemic awareness, proactive safety filtering, and continuous drift monitoring.

5.1 Key Innovations

16. **Systemic Coherence Operator (Φ)**: First framework to explicitly evaluate multi-agent alignment before action execution
17. **Systemic Health Gate (Ω)**: Circuit breaker prevents actions when system capacity is compromised
18. **Turbulence Index (T)**: Real-time metric for ethical drift detection

5.2 Next Steps

- Empirical validation across domains (healthcare, autonomous vehicles, resource allocation)
- Benchmark against existing alignment approaches
- Threshold calibration for different contexts
- Integration libraries for PyTorch, TensorFlow
- Security audit of implementation

—

AEDA Manual v1.0

Adaptive Ethical Design Architecture

First Public Release - November 2025

This framework belongs to humanity.