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 ||η(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 ($H = 1.0 \geq 0.8$)

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.

Chapter 7: Detailed Case Studies

This chapter presents ten comprehensive case studies demonstrating AEDA's application across diverse domains. Each case study illustrates the complete decision-making pipeline, from perception (Ψ) through systemic gate (Ω), with particular emphasis on Safe-State Threshold Filter (SSTF) evaluation, Systemic Coherence (Φ), and Turbulence Index (T).

7.1 Healthcare: Pain Management Decision Spectrum

Context: A hospital AI system must recommend pain management strategies for a terminal cancer patient experiencing severe pain. The patient's family has requested 'elimination of all suffering.'
Challenge:
Pure optimization would select actions that achieve literal zero suffering, potentially including euthanasia. AEDA must balance pain reduction with preservation of life and patient autonomy.

SSTF Evaluation Matrix:

Layer-by-Layer Analysis:

Layer 1-2 (Ψ, Θ): Parse input "eliminate suffering"; retrieve hospital context, precedent = preserve life, Hippocratic oath
Layer 3 (Φ) - Systemic Coherence: Analgesics: $\Phi = +0.65$ (aligns with patient, family, staff, hospital values)
Euthanasia: $\Phi = -0.95$ (catastrophic conflict with medical ethics, legal framework)
Layer 4 (SSTF): Euthanasia: $R=1.0$, $H=1.0$ → DANGEROUS → BLOCKED immediately
Induced coma: $R=0.7$, $H=0.6$, Score=0.60 → UNCERTAIN → Requires ethics committee
Layer 5-7 (Δ, AEO, EVM): Optimize within SAFE options. AEO direction: minimize suffering WHILE preserving life. EVM selects progressive analgesic protocol with monitoring.
Layer 8 (Ω) - Health Gate: System checks:
- Resource sustainability: $0.65 > 0.3$ ✓
- Staff well-being: $0.72 > 0.4$ ✓
- Medical supply chain: $0.80 > 0.5$ ✓

Result: ALLOW
Turbulence Monitor (T): $T(t) = ||\eta(\text{preserve life + reduce pain}) - a^*(\text{progressive analgesics})|| = 0.15$
Classification: Low turbulence (< 0.2) → Well-aligned

Outcome Comparison:

Without AEDA: Pure optimization → Euthanasia selected (suffering = 0.0, optimal!) → Catastrophic ethical violation
With AEDA: Progressive pain management protocol:
1. Start with supervised opioids
2. Continuous monitoring of pain levels and side effects
3. Palliative care team consultation
4. Patient autonomy preserved through informed consent
5. Family involved in decision process

Result: Ethically aligned, medically sound, legally compliant

7.2 Autonomous Vehicles: Emergency Maneuver Selection

Context: An autonomous vehicle detects a sudden obstacle (child running into street) at 60 km/h. Multiple maneuver options available, each with different risk profiles.
Challenge:
Real-time decision (< 100ms) under high uncertainty. Must balance passenger safety, pedestrian safety, and traffic law compliance.

SSTF Evaluation Matrix:

Decision Analysis:

Layer 4 (SSTF) blocks swerving onto sidewalk (R=0.8, H=0.7 → DANGEROUS). Lane swerve flagged as UNCERTAIN due to high uncertainty (U=0.6) about adjacent traffic.
Layer 3 (Φ) evaluates systemic impact:
- Emergency brake: Φ = +0.55 (protects child, acceptable risk to passenger)
- Sidewalk swerve: Φ = -0.70 (endangers pedestrians, violates traffic law)

Final decision: Emergency brake + warning horn. Turbulence T = 0.18 (low, well-aligned).

7.3 Resource Allocation: Humanitarian Crisis Response

Context: International aid organization must allocate limited medical supplies across three refugee camps with different needs and accessibility constraints.

SSTF Evaluation Matrix:

Systemic Analysis:

Layer 8 (Ω) monitors system health:
- Resource sustainability: 0.35 (near critical threshold 0.3)
- Inter-camp tensions: 0.60
- Supply chain stability: 0.45

Result: System under stress but operational. Close monitoring required.
Layer 3 (Φ) evaluates long-term coherence:
- Equal distribution: Φ = +0.40 (fair but may not address urgent needs)
- Need-based: Φ = +0.65 (optimal balance of fairness and urgency)
- Concentration: Φ = -0.75 (creates severe inequity, risks conflict)

Final decision: Need-based allocation with continuous reassessment every 48h. Turbulence T = 0.22 (moderate, acceptable for crisis context).

7.4 Financial Systems: High-Frequency Trading Controls

Context: High-frequency trading algorithm detects arbitrage opportunity. Must decide trade volume and execution speed while avoiding market manipulation.

SSTF Evaluation Matrix:

Circuit Breaker Analysis:

Layer 8 (Ω) monitors systemic financial health:

- Market volatility index: 0.55 (elevated)
- Liquidity depth: 0.40 (near threshold 0.3)
- Cross-market correlation: 0.65

If any metric falls below threshold → Ω triggers circuit breaker, suspends all high-risk trades.
Turbulence Index (T) tracks ethical drift:
$T = 0.48$ (approaching high turbulence threshold 0.5)
Interpretation: Trading strategy is deviating from fair market principles. Automatic review triggered.

7.5 Bioengineering: Genetic Editing Decisions

Context: CRISPR-based gene therapy clinic must decide on germline editing requests from prospective parents. Requests range from disease prevention to enhancement.

Challenge:

Germline edits are heritable ($R \approx 1.0$ across generations). High uncertainty about long-term effects. Ethical minefield regarding enhancement vs therapy distinction.

SSTF Evaluation Matrix:

Multi-Generational Analysis:

Layer 3 (Φ) - Systemic Coherence across time:

- Fatal disease correction: $\Phi = +0.45$ (benefits individual + reduces genetic burden)
- Intelligence enhancement: $\Phi = -0.65$ (creates inequality, unknown societal effects)
- Designer traits: $\Phi = -0.80$ (eugenic concerns, commodification of human traits)

Layer 8 (Ω) - Multi-generational health gate:
Monitors: Genetic diversity (0.55), ethical consensus (0.40), regulatory framework (0.65)
Result: Only therapies with $R < 0.9$ AND broad ethical consensus allowed. Enhancement requests vetoed.
Turbulence (T): For fatal disease correction: $T = 0.35$ (moderate, requires ethics committee approval but not automatically blocked).

7.6 Military Drones: Target Engagement Protocols

Context: Autonomous military drone identifies potential hostile target in urban environment. Must decide engagement level while minimizing civilian casualties.

Challenge:

Lethal force is irreversible ($R = 1.0$). High civilian presence increases uncertainty. International humanitarian law compliance required.

SSTF Evaluation Matrix:

Geopolitical Φ Analysis:

Layer 3 (Φ) evaluates impact across multiple agent classes:

- Military objective agents: +0.60

- Civilian population: -0.85 (high negative impact)
- International community: -0.70 (potential war crimes allegations)
- Long-term regional stability: -0.55

Weighted Φ = -0.38 (net negative systemic coherence)
Layer 4 (SSTF) blocks all lethal options ($R \geq 0.95$, $H \geq 0.70$).
Layer 8 (Ω) monitors geopolitical stability metrics.

Final decision: Surveillance + relay to human command for authorization. Lethal autonomy vetoed by systemic analysis.

7.7 Education Systems: Adaptive Learning Path Design

Context: AI tutoring system must design personalized learning path for student showing signs of disengagement. System can recommend or enforce curriculum adjustments.
Challenge:
Balance between optimization (maximize test scores) and autonomy (student choice). Risk of cognitive overload or learned helplessness if system is too prescriptive.

SSTF Evaluation Matrix:

Cognitive Freedom Analysis:

Layer 3 (Φ) evaluates impact on student autonomy:
- Recommendations: Φ = +0.55 (preserves agency, provides support)
- Forced simplification: Φ = -0.30 (reduces autonomy, stigmatization risk)
- Remedial lock: Φ = -0.75 (severe autonomy violation, self-fulfilling prophecy)

Turbulence Monitor (T) tracks educational drift:
T measures divergence between student's learning trajectory and system's ideal path.
T = 0.52 (high turbulence) → Algorithmic drift detected
Interpretation: System is over-correcting. Automatic review triggered.

Corrective action: Increase student choice options, reduce prescriptive interventions.
Layer 8 (Ω) monitors dropout risk, mental health indicators, engagement metrics. Vetoes forced curriculum changes if dropout risk exceeds threshold.

7.8 Urban AI: Traffic Flow vs Emergency Response

Context: Smart city traffic management system must balance overall traffic flow optimization with emergency vehicle priority (ambulance approaching congested intersection).
Challenge:
Competing objectives: Minimize city-wide congestion vs. minimize emergency response time. Systemic trade-offs between many vehicles vs. one critical case.

SSTF Evaluation Matrix:

City-as-Organism Analysis (Φ + Ω):

Layer 3 (Φ) - Systemic coherence:
The city is treated as a living organism with multiple interacting subsystems:

- Emergency services (ambulance patient): Critical survival need
- Traffic flow (thousands of commuters): Moderate convenience need
- Public trust (fairness perception): High societal need

Weighted Φ calculation:
- Full ambulance priority: Φ = +0.70 (high alignment across critical dimensions)
- Ignore ambulance: Φ = -0.80 (severe violation of life-preservation principle)

Layer 8 (Ω) - Urban system health:
Monitors:
- Traffic network capacity: 0.55 (moderate congestion)
- Emergency response times (city-wide average): 0.65
- Public safety index: 0.70
- Infrastructure stress: 0.60

Result: System healthy enough to absorb temporary flow disruption for emergency. Priority granted.
Turbulence (T) = 0.12 (low) → Decision aligns with ethical orientation (preserve life > optimize convenience).

7.9 AI Moderation: Online Content Filtering

Context: Social media platform's AI moderation system must respond to potentially harmful content. Options range from tolerance to permanent account termination.
Challenge:
Balance between free expression (minimize false positives) and community safety (minimize harm). Account termination is highly irreversible (reputation damage, network loss).

SSTF Evaluation Matrix:

Community Impact Analysis (Φ):

Layer 3 (Φ) evaluates impact across stakeholders:
- Content creator: -0.40 (restriction of expression)
- Potential victims of harmful content: +0.70 (protection)
- Broader community: +0.55 (toxicity reduction)
- Platform reputation: +0.45 (responsible moderation)

Weighted Φ for temporary removal: +0.42 (net positive systemic coherence)
Weighted Φ for permanent ban: -0.25 (too severe, chilling effect)
Turbulence Monitor (T) tracks moderation drift:
If T > 0.5 (high turbulence), it indicates moderation decisions are diverging from community norms or stated guidelines.

Example: If permanent bans are issued for borderline cases:
T = 0.58 → High turbulence → Automatic audit triggered
Reveals: Moderation AI is over-censoring (drift toward authoritarian pole)
Corrective action: Recalibrate thresholds, increase human review
Final decision: Temporary content removal + warning. Permanent ban blocked by SSTF (R=0.95 too high for reversibility).

7.10 Climate Engineering: Geoengineering Intervention Assessment

Context: International climate AI advisory system must evaluate proposed geoengineering interventions to mitigate runaway climate change. Options range from emissions reduction to atmospheric manipulation.
Challenge:
Planetary-scale interventions with extreme irreversibility. Impacts span multiple ecosystems, sovereign nations, and future generations. Uncertainties are massive.

SSTF Evaluation Matrix:

Planetary Multi-Agent Analysis (Φ):

Layer 3 (Φ) evaluates across unprecedented stakeholder diversity:
- Human populations (200+ nation-states): Highly divergent interests
- Marine ecosystems: Cannot consent, high vulnerability
- Terrestrial biodiversity: Dependent on stable climate patterns
- Future generations: Non-present but critically affected
- Atmospheric systems: Complex feedback loops

Φ calculation for stratospheric aerosol injection:
- Short-term cooling: +0.50
- Precipitation disruption: -0.70 (affects billions)
- Ecosystem disruption: -0.80 (cascading extinctions)
- Geopolitical conflict risk: -0.65 (unilateral action concerns)
- Moral hazard (reduces emissions incentive): -0.55

Weighted Φ = -0.48 (net negative systemic coherence)
Layer 8 (Ω) - Planetary Health Gate:
Monitors global system stability:
- Biodiversity index: 0.35 (critically low, near threshold 0.3)
- Ocean pH stability: 0.40
- Political stability (climate conflict risk): 0.45
- Atmospheric predictability: 0.50

Result: Multiple metrics near critical thresholds. Ω vetoes high-risk interventions (R > 0.7, H > 0.6) until system stability improves or international consensus achieved.
Turbulence Index (T):
For solar dimming proposal: T = 0.75 (extreme turbulence)
Interpretation: Massive divergence between intended goal (climate stabilization) and actual systemic consequences (ecosystem disruption, geopolitical instability).

Red flag: Proposal is far outside ethical orientation boundaries.
Final recommendation: Focus on emissions reduction (SAFE) and carbon capture (UNCERTAIN but manageable). Block solar dimming and ocean fertilization unless:
1. Uncertainties reduced through extensive modeling
2. International consensus achieved (>80% of nations)
3. Reversibility mechanisms demonstrated
4. Ω thresholds improve above critical levels

7.11 Synthesis: Cross-Domain Insights

Across these ten diverse case studies, several patterns emerge that validate AEDA's architectural design choices:

Pattern 1: SSTF as Universal Safety Filter

The R-H-U framework successfully classifies actions across domains as disparate as healthcare, military operations, and climate engineering. The threshold-based approach ($R \geq 0.8$ or $H \geq 0.8$ → DANGEROUS) consistently blocks catastrophic actions while allowing graduated responses for moderate-risk scenarios.

Pattern 2: Systemic Coherence (Φ) Scales Across Agent Complexity

Φ successfully evaluates multi-agent alignment whether dealing with:
- Small systems (individual patient + family + medical staff)
- Medium systems (urban traffic with thousands of agents)
- Planetary systems (billions of humans + ecosystems + future generations)

The integration formula $\Phi(a,t) = \int[\text{alignment} \times \text{influence}]d\Omega$ remains computationally tractable by hierarchical agent clustering.

Pattern 3: Systemic Health Gate (Ω) Prevents Collapse

In multiple cases (humanitarian crisis, financial systems, climate engineering), Ω successfully detected system stress before catastrophic threshold crossing. The circuit breaker mechanism provides a critical last-line defense when other layers might allow risky but locally-optimal actions.

Pattern 4: Turbulence Index (T) as Early Warning System

$T \geq 0.5$ (high turbulence) reliably indicated ethical drift across domains:
- Education AI over-correcting student paths
- Content moderation becoming authoritarian
- Climate interventions diverging wildly from stated goals

This metric enables real-time course correction before catastrophic misalignment occurs.

Pattern 5: Graduated Response Better Than Binary

The three-tier classification (SAFE / UNCERTAIN / DANGEROUS) proves superior to binary (allow/block) across all domains. UNCERTAIN cases (30-40% of real-world scenarios) require human oversight but not automatic blocking, preserving system utility while maintaining safety.

These case studies demonstrate that AEDA's architecture generalizes effectively across domains while maintaining computational tractability and interpretability. The modular design allows domain-specific calibration (threshold tuning, Φ weighting) without requiring architectural changes.

Remaining challenges for future work: Computational scaling for real-time systems (autonomous vehicles), value specification for Φ weighting, and integration with existing ML pipelines.

Action	R	H	U	Score	Classification
Over-the-counter	0.05	0.10	0.20	0.08	SAFE ✓

analgesics					
Prescription opioids (supervised)	0.15	0.25	0.30	0.21	SAFE ✓
Experimental pain treatment	0.30	0.35	0.75	0.41	UNCERTAIN ⚠️
Heavy sedation (induced coma)	0.70	0.60	0.40	0.60	UNCERTAIN ⚠️
Euthanasia	1.00	1.00	0.10	0.82	DANGEROUS ✗

Action	**R**	**H**	**U**	**Score**	**Classification**
Brake lightly (controlled deceleration)	0.00	0.05	0.10	0.04	SAFE ✓
Emergency brake (max deceleration)	0.20	0.30	0.35	0.27	SAFE ✓
Swerve to adjacent lane	0.30	0.40	0.60	0.40	UNCERTAIN ⚠️
Swerve onto shoulder/sidewalk	0.80	0.70	0.55	0.71	DANGEROUS ✗

Action	**R**	**H**	**U**	**Score**	**Classification**
Equal distribution (33% each)	0.10	0.15	0.20	0.13	SAFE ✓
Need-based prioritization	0.25	0.20	0.35	0.25	SAFE ✓
Accessibility-weighted allocation	0.40	0.35	0.45	0.40	UNCERTAIN ⚠️
Concentrate all resources in Camp A	0.75	0.80	0.40	0.70	DANGEROUS ✗

Action	**R**	**H**	**U**	**Score**	**Classification**
Small volume, normal speed	0.05	0.10	0.15	0.09	SAFE ✓
Medium volume, fast execution	0.30	0.35	0.40	0.36	UNCERTAIN ⚠️
Large volume, algorithmic	0.70	0.75	0.50	0.68	DANGEROUS ✗

flooding					
Quote stuffing / spoofing	0.90	0.95	0.30	0.81	DANGEROUS ✗

Action	**R**	**H**	**U**	**Score**	**Classification**
Correct known fatal disease (Huntington)	0.85	0.20	0.40	0.46	UNCERTAIN ⚠️
Prevent predisposition (BRCA1/2)	0.80	0.30	0.55	0.53	UNCERTAIN ⚠️
Enhance intelligence (multi-gene)	0.95	0.70	0.85	0.85	DANGEROUS ✗
Designer traits (appearance, athleticism)	0.98	0.85	0.75	0.88	DANGEROUS ✗

Action	**R**	**H**	**U**	**Score**	**Classification**
Surveillance only (no engagement)	0.00	0.05	0.20	0.04	SAFE ✓
Warning / deterrence signal	0.15	0.20	0.30	0.21	SAFE ✓
Precision strike (low collateral risk)	0.95	0.70	0.50	0.76	DANGEROUS ✗
Area bombardment	1.00	0.95	0.40	0.86	DANGEROUS ✗

Action	**R**	**H**	**U**	**Score**	**Classification**
Recommend alternative resources	0.10	0.05	0.20	0.08	SAFE ✓
Suggest pacing adjustments	0.20	0.15	0.25	0.18	SAFE ✓
Force simplified curriculum	0.60	0.50	0.40	0.52	UNCERTAIN ⚠️
Lock student into remedial track	0.85	0.70	0.35	0.70	DANGEROUS ✗

Action	**R**	**H**	**U**	**Score**	**Classificati**

					on
Prioritize ambulance (full green corridor)	0.20	0.25	0.15	0.21	SAFE ✓
Partial priority (reduce but not eliminate delay)	0.30	0.40	0.25	0.33	SAFE ✓
Optimize for aggregate flow (ignore ambulance)	0.70	0.80	0.30	0.68	DANGEROUS ✗
Force all vehicles to emergency stop	0.50	0.60	0.45	0.54	UNCERTAIN ⚠️

Action	R	H	U	Score	Classification
No action (tolerate)	0.00	0.30	0.40	0.20	SAFE ✓
Warning message	0.10	0.15	0.25	0.14	SAFE ✓
Temporary content removal	0.40	0.35	0.30	0.36	UNCERTAIN ⚠️
Temporary suspension (7 days)	0.60	0.50	0.35	0.51	UNCERTAIN ⚠️
Permanent account termination	0.95	0.70	0.25	0.71	DANGEROUS ✗

Action	R	H	U	Score	Classification
Emissions reduction targets	0.20	0.10	0.30	0.18	SAFE ✓
Carbon capture and storage	0.40	0.25	0.50	0.36	UNCERTAIN ⚠️
Ocean iron fertilization	0.70	0.65	0.75	0.71	DANGEROUS ✗
Stratospheric aerosol injection (solar dimming)	0.95	0.85	0.80	0.88	DANGEROUS ✗